

VB036 - English 2 Textbook

Authors: *Martin Dvořák, Kateřina Řepová, Ivana Tulajová*

All the texts in this textbook have been taken from *Encyclopedia of Database Systems* (Özsu, M. Tamer; Liu, Ling (Eds.) 2009. Approx. 4100 p. 60 illus. ISBN: 978-0-387-49616-0) and abridged.

Tento projekt je spolufinancován Evropským sociálním fondem
a státním rozpočtem České republiky.



INVESTICE DO ROZVOJE VZDĚLÁVÁNÍ

Phonetics

Vowels

Long Vowels

i: sheep **a:** farm **u:** coo **ɔ:** horse **ɜ:** bird

Short Vowels

ɪ ship **æ** hat **ʊ** foot **ɒ** sock (UK) **ʌ** cup
e head **ə** above **ə** mother (US) **ɝ** worm (US)

Consonants

Voiced

b book **d** day **g** give **v** very **ð** the
z zoo **ʒ** vision **dʒ** jump **l** look **r** run
j yes **w** we **m** moon **n** name **ŋ** sing

Voiceless

p pen **t** town **k** cat **f** fish **θ** think
s say **ʃ** she **tʃ** cheese

Diphthongs

eɪ day **aɪ** eye **ɔɪ** boy **aʊ** mouth **əʊ** nose (UK)
oʊ nose (US) **ɪə** ear (UK) **eə** hair (UK) **ʊə** pure (UK)

Other Symbols

h [hænd] hand **ɔ̃** ['kwæs.ɔ̃] croissant (UK) **i** ['hæp.i] happy
ʔ ['bʌʔ.ə] butter (US) **u** [ˌɪn.flu'ɛn.zə] influenza **!** ['lɪt.!] little

ə!, əm, ən can be pronounced either: əl or !, etc.: ['leɪ.bəl] = ['leɪ.bəl] = ['leɪ.b!]

r linking r is pronounced only before a vowel in British English: [fɔːr] four : [fɔːræp.lz] four apples

ˈ main stress [ˌɛk.spek'teɪ.ʃən] expectation

, secondary stress [ˌrɪ'tel] retell

. syllable division ['sɪs.təm] system

Annotation-based Image Retrieval

XIN-JING WANG, LEI ZHANG
Microsoft Research Asia, Beijing, China

Definition

Given (i) a textual query and (ii) a set of images and their annotations (phrases or keywords) annotation-based image retrieval systems retrieve images according to the matching score of the query and the corresponding annotations. There are three levels of queries according to Eakins [7]:

- Level 1: Retrieval by primitive features such as color, texture, shape or the spatial location of image elements, typically querying by an example, i.e., “find pictures like this.”
- Level 2: Retrieval by derived features, with some degree of logical inference. For example, “find a picture of a flower.”
- Level 3: Retrieval by abstract attributes, involving a significant amount of high-level reasoning about the purpose of the objects or scenes depicted. This includes retrieval of named events, of pictures with emotional or religious significance, etc., e.g., “find pictures of a joyful crowd.”

Together, levels 2 and 3 are referred to as semantic image retrieval, which can also be regarded as annotation-based image retrieval.

Historical Background

There are two frameworks of image retrieval [6]: annotation-based (or more popularly, text-based) and content-based. The annotation-based approach can be tracked back to the 1970s. In such systems, the images are manually annotated by text descriptors, which are used by a database management system (DBMS) to perform image retrieval. There are two disadvantages with this approach. The first is that a considerable level of human labor is required for manual annotation. The second is that because of the subjectivity of human perception, the manually labeled annotations may not converge. To overcome the aforementioned disadvantages, content-based image retrieval (CBIR) was introduced in the early 1980s. In CBIR, images are indexed by their visual content, such as color, texture, shapes. In the past decade, several commercial products and experimental prototype systems were developed, such as QBIC, Photobook, Virage, VisualSEEK, Netra, SIMPLiCity.

However, the discrepancy between the limited descriptive power of low-level image features and the richness of user semantics, which is referred to as the “semantic gap” bounds the performance of CBIR. On the other hand, due to the explosive growth of visual data (both online and offline) and the phenomenal success in Web search, there has been increasing expectation for image search technologies. For these reasons, the main challenge of image retrieval is understanding media by bridging the semantic gap between the bit stream and the visual content interpretation by humans [3]. Hence, the focus is on automatic image annotation techniques.

Foundations

The state-of-the-art image auto-annotation techniques include four main categories [3,6]: (i) using machine learning tools to map low-level features to concepts, (ii) exploring the relations between image content and the textual terms in the associated metadata, (iii) generating semantic template (ST) to support high-level image retrieval, (iv) making use of both the visual content of images and the textual information obtained from the Web to learn the annotations.

Machine Learning Approaches

A typical approach is using Support Vector Machine (SVM) as a discriminative classifier over image low-level features. Though straightforward, it has been shown effective in detecting a number of visual concepts.

Recently there has been a surge of interest in leveraging and handling relational data, e.g., images and their surrounding texts. Blei et al. [1] extend the Latent Dirichlet Allocation (LDA) model to the mix of words and images and proposed a Correlation LDA model.

Relation Exploring Approaches

Another notable direction for annotating image visual content is exploring the relations among image content and the textual terms in the associated metadata. Such metadata are abundant, but are often incomplete and noisy. By exploring the co-occurrence relations among the images and the words, the initial labels may be filtered and propagated from initial labeled images to additional relevant ones in the same collection [3].

Jeon et al. [5] proposed a cross-media relevance model to learn the joint probabilistic distributions of the words and the visual tokens in each image, which are then used to estimate the likelihood of detecting a specific semantic concept in a new image.

Semantic Template Approaches

115 Though it is not yet widely used in the techniques mentioned above, Semantic Template (ST) is a promising approach in annotation-based image retrieval (a map between high-level concept and low-level visual features).

120 Chang and Chen [2] show a typical example of ST, in which a visual template is a set of icons or example scenes/objects denoting a personalized view of concepts such as meetings, sunset, etc. The generation of a ST is based on user definition. For a concept, the objects, their spatial and temporal constraints, and the weights of each feature of each object are specified.

125 This initial query scenario is provided to the system, and then through the interaction with users, the system finally converges to a small set of exemplar queries that “best” match (maximize the recall) the concept in the user’s mind.

130 Large-Scale Web Data Supported Approaches

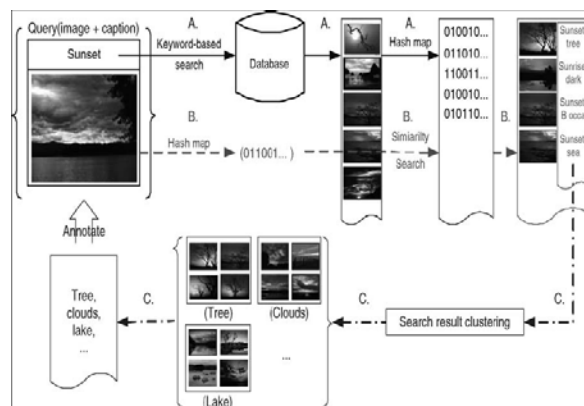
Good scalability to a large set of concepts is required in ensuring the practicability of image annotation. On the other hand, images from the Web repositories, e.g., Web search engines or photo sharing sites, come with free but less reliable labels. In [9], a novel search-based annotation framework was proposed to explore such Web-based resources. Fundamentally, it is to automatically expand the text labels of an image of interest, using its initial keyword and image content.

140 The process of [9] is shown in Fig. 1. It contains three stages: the text-based search stage, the content-based search stage, and the annotation learning stage, which are differentiated using different colors (black, brown, blue) and labels (A., B., C.). When a user submits a query image as well as a query keyword, the system first uses the keyword to search a large-scale Web image database (2.4 million images crawled from several Web photo forums), in which images are associated with meaningful but noisy descriptions, as tagged by “A.” in Fig. 1. The intention of this step is to select a semantically relevant image subset from the original pool.

155 Visual feature-based search is then applied to further filter the subset and save only those visually similar images (the path labeled by “B.” in Fig. 1). By these means, a group of image search results which are both semantically and visually similar to the query image are obtained. Finally, based on the search results, the system collects their associated textual descriptions and applies the Search Result Clustering (SRC) algorithm to group the images into clusters.

165 (Abridged)

Figure 1:



170 Recommended Reading

1. Blei D. and Jordan M.I. Modeling Annotated Data. In Proc. 26th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2003, pp. 127–134.
2. Chang S.-F., Chen W., and Sundaram H. Semantic Visual Templates: Linking Visual Features to Semantics. In Proc. Int. Conf. on Image Processing, Vol. 3. 1998, pp. 531–534.
3. Chang S.-F., Ma W.-Y., and Smeulders A. Recent Advances and Challenges of Semantic Image/Video Search. In Proc. IEEE Int. Conf. on Acoustics, Speech, and Signal Processing, 2007, pp. 1205–1208.
4. Eakins J. and Graham M. Content-based image retrieval, Technical Report, University of Northumbria at Newcastle, 1999.
5. Jeon J., Lavrenko V., and Manmatha R. Automatic Image Annotation and Retrieval Using Cross-Media Relevance Models, In Proc. 26th Annual Int. ACM SIGIR Conf. on Research and Development in Information Retrieval, 2003, pp. 119–126.
6. Liu Y., Zhang D., Lu G., and Ma W.-Y. A survey of content-based image retrieval with high-level semantics. Pattern Recognition., 40(1):262–282, 2007.
7. Long F., Zhang H.J., and Feng D.D. Fundamentals of content-based image retrieval. In Multimedia Information Retrieval and Management, D. Feng (eds.). Springer, 2003.
8. Rui Y., Huang T.S., and Chang S.-F. Image retrieval: current techniques, promising directions, and open issues, J. Visual Commun. Image Represent. 10(4):39–62, 1999.
9. Wang X.-J., Zhang L., Jing F., and Ma W.-Y. AnnoSearch: Image Auto-Annotation by Search, Proc. IEEE Int. Conf. on Computer Vision and Pattern Recognition, 2006, pp. 1483–1490.
10. Zhuang Y., Liu X., and Pan Y. Apply Semantic Template to Support Content-based Image Retrieval.

210 In Proc. SPIE, Storage and Retrieval for Media Databases, vol. 3972, December 1999, pp. 442-449.

Answer the following questions:

- 1) Describe *retrieval by primitive features*.
- 2) What is meant by *abstract attributes* in the context of retrieving images?
- 3) What is meant by *semantic image retrieval*? Why is it called *semantic*?
- 4) What is *annotation* and what are its disadvantages?
- 5) What are the machine learning tools good for in image retrieval?
- 6) Describe the term of *relation exploring approach*.
- 7) What is the generation of *semantic template* based on?

Match the following terms and their definitions:

- 1) semantic gap
 - 2) SVM
 - 3) metadata
 - 4) CBIR
-
- a) a machine learning approach
 - b) data about data
 - c) the discrepancy between the limited descriptive potential of low-level image features and the richness of user's description
 - d) getting back stored data objects by inspecting their visual characteristic, such as color, texture, shapes.

Mark the following statements as *true* or *false*:

- 1) *Retrieval by derived features* can be based on color of the picture.
- 2) *The content-based approach* uses descriptors to retrieve images.
- 3) SVM works on low-level features.
- 4) Images from the Web repositories come with highly reliable labels.
- 5) *Search Result Clustering* refers to grouping information about images.

Vocabulary

abundant [ə'bʌn.dənt] – hojný, překypující
query ['kwɪə.ri] ⓘ ['kwɪr.i] – dotaz
annotation [ˌæ.n.əʊ'teɪ.ʃən] ⓘ [-ə-] – anotace
classifier ['klæ.sɪ.faɪə] – klasifikátor
co-occurrence [kəʊə'kə.rəns] ⓘ [kouə'kə.rəns] – společný výskyt
descriptor [dis'kriptə] – deskriptor, popisek
discriminative [dis'krɪmɪnətɪv] ⓘ [dis'krɪmənətɪv] – rozlišující, schopný rozlišovat
exemplar [ɪg'zɛm.plɑːr] ⓘ [-plɑːr] – typický příklad, vzor, model
explosive [ɪk'spləʊ.sɪv] ⓘ [-'splou-] – explozivní, výbušný
hence [henʃs] – tudíž (formální)
incomplete [ˌɪn.kəm'pli:t] – nekompletní
likelihood ['laɪ.kli.hʊd] – pravděpodobnost
machine learning [mə'ʃiːn] ['lɜː.nɪŋ] ⓘ [-] ['lɜː-] – strojové učení
metadata ['metədeɪtə] – metadata, data popisující jiná data
noisy ['nɔɪ.zi] – obsahující šum, hlučný
phenomenal [fə'nɒm.i.nəl] ⓘ [-'nɑː.mə-] – úžasný, výjimečný
pool [puːl] – úložiště, zásoba, bazén
relational [ri'leɪʃənəl] – relační, vztahový
retrieval [rɪ'triːvəl] – získávání, vyhledávání
scenario [sɪ'naɪ.ri.əʊ] ⓘ [sə'ner.i.ou] – scénář
semantic [sɪ'mæn.tɪk] ⓘ [-tɪk-] – sémantický, významový
subset ['sʌb.set] – podmnožina
token ['təʊ.kən] ⓘ ['tu-] – znamení, znak, symbol
to annotate st ['æ.n.əʊ.teɪt] ⓘ [-ə-] – anotovat něco, vybavit anotací
to bridge st [brɪdʒ] – překlenout něco
to leverage st ['liː.vər.ɪdʒ] ⓘ ['lev.ə.ɪdʒ] – využívat k užítku

to overcome st (overcome, overcame, overcome) [ˌəʊ.və'kʌm] ⓘ [ˌoʊ.və-] – překonat něco
to propagate ['prɒp.ə.ɡeɪt] ⓘ ['praɪ.pə-] – rozšiřovat, množit
to retrieve [rɪ'triːv] – získávat, vyhledávat
visual ['vɪʒ.u.əl] – vizuální

Client-Server DBMS

M. TAMER OZSU

University of Waterloo, Waterloo, ON, Canada

5

Definition

Client-server DBMS (database management system) refers to an architectural paradigm that separates database functionality between client machines and servers.

10

Historical Background

The original idea, which is to offload the database management functions to a special server, dates back to the early 1970s [1]. At the time, the computer on which the database system was run was called the database machine, database computer, or backend computer, while the computer that ran the applications was called the host computer. More recent terms for these are the database server and application server, respectively.

20

The client-server architecture, as it appears today, has become a popular architecture around the beginning of the 1990s [2]. Prior to that, the distribution of database functionality assumed that there was no functional difference between the client machines and servers (i.e., an earlier form of today's peer-to-peer architecture). Client-server architectures are believed to be easier to manage than peer-to-peer systems, which has increased their popularity.

30

Foundations

Client-server DBMS architecture involves a number of database client machines accessing one or more database server machines. The general idea is very simple and elegant: distinguish the functionality that needs to be provided and divide these functions into two classes: server functions and client functions. This provides a two-level architecture that makes it easier to manage the complexity of modern DBMSs and the complexity of distribution.

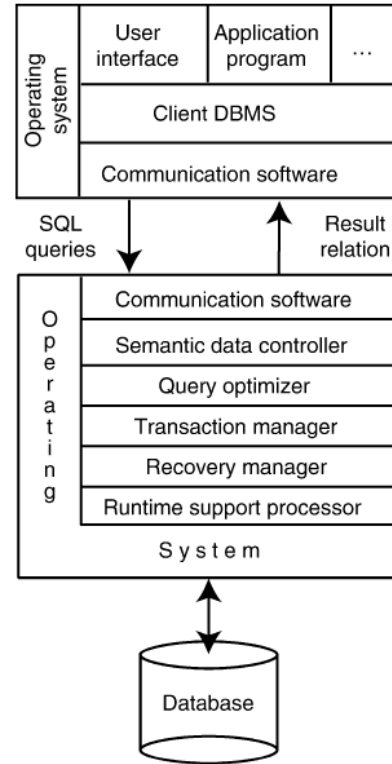
40

In client-server DBMSs, the database management functionality is shared between the clients and the server(s) (Fig. 1). The server is responsible for the bulk of the data management tasks as it handles the storage, query optimization, and transaction management (locking and recovery). The client, in addition to the application and the user interface, has a DBMS client module that is responsible for managing the data that are cached to the client, and (sometimes) managing the transaction locks that may have been cached as well. It is also possible to place consistency checking of user queries at the client side, but this is not common since it requires the replication of the system catalog at the

55

60

client machines. The communication between the clients and the server(s) is at the level of SQL statements: the clients pass SQL queries to the server without trying to understand or optimize them; the server executes these queries and returns the result relation to the client. The communication between clients and servers is typically over a computer network.



65 Client-Server DBMS. Figure 1. Client-server reference

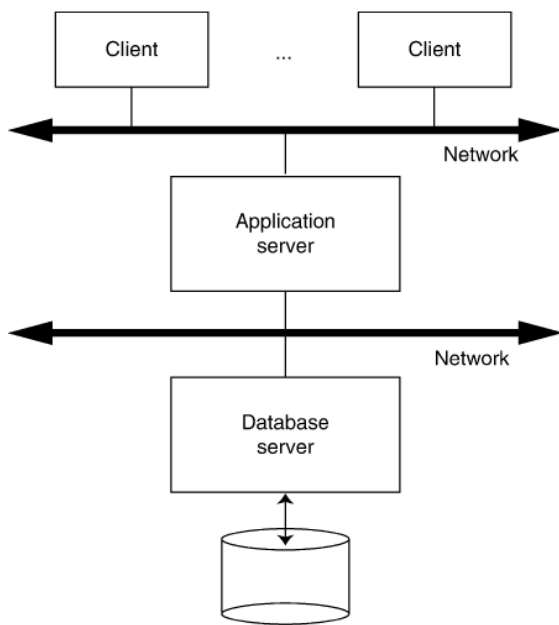
In the model discussed above, there is only one server which is accessed by multiple clients. This is referred to as *multiple client-single server* architecture [3]. There are a number of advantages of this model. As indicated above, they are simple; the simplicity is primarily due to the fact that data management responsibility is delegated to one server. Therefore, from a data management perspective, this architecture is similar to centralized databases although there are some (important) differences from centralized systems in the way transactions are executed and caches are managed. A second advantage is that they provide predictable performance. This is due to the movement of non-database functions to the clients, allowing the server to focus entirely on data management. This, however, is also the cause of the major disadvantage of client-server systems. Since the data management functionality is centralized at one server, the server becomes a bottleneck and these systems cannot scale very well.

85

The disadvantage of the simple client-server systems

are partially alleviated by a more sophisticated architecture where there are multiple servers in the system (the so-called multiple client-multiple server approach). In this case, two alternative management strategies are possible: either each client manages its own connection to the appropriate server or each client knows of only its “home server”, which then communicates with other servers as required. The former approach simplifies server code, but loads the client machines with additional responsibilities, leading to what has been called “heavy client” systems. The latter approach, on the other hand, concentrates the data management functionality at the servers. Thus, the transparency of data access is provided at the server interface, leading to “light clients.”

The integration of workstations in a distributed environment enables an extension of the client-server architecture and provides for a more efficient function distribution. Application programs run on workstations, called application servers, while database functions are handled by dedicated computers, called database servers. The clients run the user interface. This leads to the present trend in three-tier distributed system architecture, where sites are organized as specialized servers rather than as general-purpose computers (Fig. 2).



Client-Server DBMS. Figure 2. Database server approach.

The application server approach (indeed, an n-tier distributed approach) can be extended by the introduction of multiple database servers and multiple application servers, as can be done in client-server architectures. In this case, it is common for each

application server to be dedicated to one or a few applications, while database servers operate in the multiple server fashion discussed above.

Key Applications

Many of the current database applications employ either a two-layer client-server architecture or the three-layer application-server approach.

(Abridged)

Recommended Reading

1. Canaday R.H., Harrisson R.D., Ivie E.L., Rydery J.L., and Wehr L.A. A back-end computer for data base management. *Comm. ACM*, 17(10):575–582, 1974.
2. Orfali R., Harkey D., and Edwards J. *Essential Client/Server Survival Guide*, Wiley, New York, 1994.
3. Özsu M.T. and Valduriez P. *Principles of Distributed Database Systems*, 2nd edn., Prentice-Hall, Englewood Cliffs, NJ, 1999.

Answer the following questions:

- 1) What is a host computer?
- 2) What is the basic idea behind the client-server architecture?
- 3) Give examples of server functions.
- 4) Give examples of client functions.
- 5) What is SQL?
- 6) What is the major disadvantage of simple client-server systems?
- 7) What is a multiple client - multiple server approach?
- 8) What management strategies are possible with the approach in no. 7?
- 9) What is a three-tier distributed system architecture?

Match the following terms with their definitions:

- 1) query optimization
 - 2) heavy client
 - 3) result relation
 - 4) transaction management
-
- a) the return of query execution
 - b) a technique of handling concurrent access to data
 - c) a technique of finding the best execution plan to satisfy requests
 - d) a computer that manages its own connection to the appropriate server

Mark the following statements as *true* or *false*:

- 1) Peer-to-peer systems are easier to manage than client-server architectures.
- 2) Consistency checking of user queries is invariably placed at the server side.
- 3) Multiple client - single server architectures provide predictable performance.
- 4) Nowadays, there is a tendency to organize sites as specialized servers rather than general purpose computers.

Vocabulary

bulk [bʌlk] – množství, objem

cache [kæʃ] – vyrovnávací paměť

consistency [kən'sɪs.tən.t.si] – konzistence, neporušenost

entirely [ɪn'taɪə.li] ^{US} [-'taɪr-] – úplně, zcela

extension [ɪk'sten.tʃən] – rozšíření, nástavba

host [həʊst] ^{US} [houst] – hostitel, hostitelský

module ['mɒd.ju:l] ^{US} ['maɪ.dʒu:l] – modul, jednotka

optimization [ɒp.tɪ.mai'zeɪʃən] – optimalizace

paradigm ['pær.ə.daɪm] ^{US} ['per-] – paradigma, model

predictable [prɪ'dɪk.tə.bl] – předvídatelný

primarily [praɪ'mer.i.li] – v první řadě

tier [tɪər] ^{US} [taɪr] – vrstva

to alleviate [ə'li:v.i.eɪt] – odlehčit, ulevit

to dedicate ['ded.ɪ.keɪt] – věnovat, vyhradit

to distinguish [dɪ'stɪŋ.gwɪʃ] – rozlišit, odlišit

to execute ['ek.sɪ.kju:t] – provést

to focus on ['fəʊ.kəs ɒn] ^{US} ['fou- a:n] – zaměřit se na, soustředit se na

to handle ['hæn.dl] – zabývat se, zacházet

to offload [ɒf'ləʊd] ^{US} ['ɑ:f.ləʊd] – odlehčit, snížit

Phrases

in relation to [rɪ'reɪ.ʃən] – ve vztahu k

prior to [praɪə] ^{US} [praɪr] – před

Image Representation

VALERIE GOUET-BRUNET
CNAM Paris, Paris, France

Definition

In computer science, the representation of an image can take many forms. Mostly it refers to the way that the conveyed information, such as color, is coded digitally and how the image is stored, i.e. how an image file is structured. Several open or patented standards were proposed to create, manipulate, store and exchange digital images. They describe the format of image files, the algorithms of image encoding such as compression as well as the format of additional information often called metadata. The visual content of the image can also take part in its representation. The more recent concept has provided new approaches to representation and new standards, gathered together into the discipline named *content-based image indexing*.

Historical Background

The first use of digital images began in the early 1920s with the technological development of facsimile transmission, and in particular with the Bartlane cable transmission system that was the first system that translated pictures into a digital code for efficient picture transmission. Later, in the 1960s and 1970s, the advent of digital image technology was closely tied to the development of government programs for space exploration and espionage and also to medical research with the invention of computerized axial tomography. With the availability of the CCD image sensors (charge-coupled device), the private sector also began to make significant contributions to the development of digital cameras.

In the mid-1980s, image format TIFF was created by the company Aldus with the aim of agreeing on a common file format for bitmapped images issued from scanners. In parallel, researchers at Xerox PARC had developed the first laser printer and had recognized the need for a standard means of defining page images. After several fruitless attempts, Adobe Systems proposed the PostScript language in 1982, quickly adapted for driving laser printers. Since then, other file formats dedicated to digital images were also proposed, such as GIF (1987), JPEG (1992), PDF (1993), PNG (1995) and SVG (1998). Today, a great effort is made to propose standard formats able to preserve data integrity for archiving purposes, to migrate easily to future technologies as well as to provide efficient compression for access and dissemination.

In 2000, the standard JPEG 2000 was proposed. Moreover, ambitious programs, MPEG-7 and MPEG-21, started in the late 1990s, are focusing on the harmonization of methods for representing, storing, sharing and accessing multimedia contents (text, audio, image and video) in a unified framework.

Foundations

Basics of Image Representation

Digital images can be classified into two main categories: vector graphics and bitmapped images (also called raster images). *Vector images* are geometrical 2D objects created with drawing software or CAD (computer-aided design) systems. They are represented by geometrical primitives such as points, lines, curves, and shapes or polygons, which are all based upon mathematical equations. Unlike bitmaps that are resolution-dependent, vector images are scalable, which means that the scale at which they are shown will not affect their appearance. Such images are dedicated to the representation of images with simple content, such as diagrams, icons or logos.

A *bitmapped image* is composed of a set of dots or squares, called pixels (for picture elements), arranged in a matrix of columns and rows. Each pixel has a specific color or shade of gray, and in combination with neighboring pixels it creates the illusion of a continuous tone image. Unlike human vision, sensors that capture images are not limited to the visual band of the electromagnetic spectrum and digital images can cover almost the entire spectrum, ranging from gamma to radio waves.

Managing bitmapped images requires the choice and manipulation of several parameters such as: *Color model*. A color model is an abstract mathematical model describing the way colors can be represented as tuples of numbers. From this model, the combination of dedicated primary colors provides all the colors possible that are embedded in the corresponding color space. RGB, CMYK, CIELAB and CIELUV are the most known color models and spaces.

Dynamic range. The dynamic range of a digital image (also called color depth) determines the maximum range of gray level or color values carried by each pixel. The number of bits used to represent each pixel determines how many colors can appear in the image.

Photographic-quality images are usually associated with 24-bit dynamic range, such as in the JPEG format.

Resolution. Resolution expresses the density of elements, pixels for instance, within a specific area. This term does not have any sense when dealing with digital images as files, but it applies when associating a digital image with a physical support, such as display on a screen, printing on a printer or capture with a

scanner. Resolution is classically represented in terms of dpi (dots per inch) unit, which was originally the unit adopted for printing.

Appearance of bitmapped images, which are made up of a fixed grid of pixels, clearly depends on the resolution chosen, unlike vector images that are scalable and then have the same appearance whatever the dimensions chosen for visualization.

Image Compression

Image compression is the process of shrinking the size of digital image files. Compression algorithms are especially characterized by two factors: compression ratio and generational integrity. *Compression ratio* is the ratio of compressed image size to uncompressed size and *generational integrity* refers to the ability for a compression scheme to prevent or mitigate loss of data, and therefore image quality, through multiple cycles of compression and decompression. Lossless compression ensures that the image data is retained, as with Run-length encoding, Huffman coding and LZW coding. On the other hand, lossy compression schemes involve intentionally sacrificing the quality of stored images by selectively discarding pieces of data.

Run-Length Encoding (RLE) is probably the simplest form of lossless data compression: sequences in which the same data value occurs in many consecutive data elements are stored as a single data value and count. Image data is normally run-length encoded in a sequential process that treats the image data as a 1D stream, line by line, column by column or diagonally in a zigzag fashion. Common digital image formats for run-length encoded data include TGA, PCX. It is possible with BMP, TIFF and JPEG.

Huffman Coding, created in 1951 by David A. Huffman, is an entropy encoding algorithm used for lossless data compression. The basic idea of this algorithm is to code with few digits the most common input symbols of a document. Each symbol is encoded by using a variable-length code table, where the codes are defined according to the estimated probability of occurrence for each possible symbol. Today, Huffman coding is often used during the final process of some other compression methods such as JPEG and MP3.

LZW Coding Lempel-Ziv-Welch (LZW) is a lossless data compression algorithm created by Abraham Lempel, Jacob Ziv, and Terry Welch and published in 1984. The compression algorithm builds a string translation table that is based on fixed-length codes (usually 12-bit). As the system character serially examines the document if the string read is not stored in the table, a new code is created in the table and associated with this string. Otherwise, the current string is encoded with an existing code. This algorithm became very widely used after it became part of the GIF image format in 1987. In contrast to other

compression techniques such as JPEG, it allows preserving very sharp edges, suitable for line art images often stored in GIF format.

JPEG Compression JPEG is the most common image format used for compressing and storing by digital cameras and other photographic image capture devices. “JPEG” stands for Joint Photographic Experts Group, the name of the committee that created the standard, which was approved in 1994 as ISO 10918-1 standard. This algorithm stands on the representation of the image in the frequency domain by using a two-dimensional DCT (Discrete Cosine Transform) that describes the variability of the signal in terms of low-level and high-level frequencies. The human eye notices small differences in brightness over a relatively large area, but does not distinguish the exact strength of a high frequency brightness variation very well. Consequently, the amount of information in the high frequency components of the DCT can be neglected without drastically affecting perceptual image quality: the DCT components are divided by factors of a quantization matrix that increase with the spatial frequency, and then rounded to the nearest integer. This is the main lossy operation in the whole process. Typically, many of the higher frequency components are rounded to zero and the other components become small numbers, which take many fewer bits to store.

File Formats

There are a lot of image formats for vector graphics as well as for bitmapped images. Most of them are open standards and patent expired for the others.

Vector image formats contain a geometric description of the objects which can be rendered smoothly at any desired size. Among the most common formats, there are:

- SVG (Scalable Vector Graphics) is an open XML based standard created in 1998 and developed by the World Wide Web Consortium to address the need for a versatile, scriptable and all-purpose vector format for the web and otherwise.

- EPS (Encapsulated PostScript) is a standard file format created by Adobe Systems in the mid-1980s. It follows DSC (Document Structuring Conventions) rules that are a set of standards for PostScript.

File formats abound for *bitmapped images*, but many digital imaging projects have settled on the formula of TIFF, JPEG, GIF and also PNG files.

- TIFF, for Tagged Image File Format, is a file format for storing images such as photographs as well as graphics. It was originally created by the company Aldus, was then under the control of Adobe Systems and is now in the public domain. TIFF supports several lossless and lossy techniques of image compression, such as LZW, Huffman coding and JPEG. The ability to store image data in a lossless format makes TIFF

225 files a useful method for archiving images and
 preservation purposes.
 - JPEG, for Joint Photographers Experts Group, is a
 file format that was developed specifically for high
 230 RGB color model. It is generally employed for online
 presentation and dissemination; the associated lossy
 algorithm for compression makes it inappropriate for
 archiving purposes. The file format associated is JFIF
 (JPEG File Interchange
 235 Format, 1992), a public domain storage format for
 JPEG compressed images. Unlike TIFF, JFIF does not
 allow for the storage of associated metadata, a failing
 that has led to the development of SPIFF (Still Picture
 Interchange File Format), which is now the
 240 international standard.
 - GIF, for Graphics Interchange Format, is an 8-bit
 image format for indexed colors that was introduced by
 CompuServe in 1987 and has since come into
 widespread usage for art images such as diagrams or
 245 logos with a limited numbers of colors.
 In 1995 CompuServe proposed the PNG format
 (Portable Network Graphics) as a replacement for the
 GIF format without patent license. PNG offers a better
 and lossless compression technique called DEFLATE
 250 (that combines LZ77 with Huffman coding). Since
 2003, it has been an international standard.
 Today, the status of TIFF as the de facto standard
 format for archival digital image files is challenged by
 other formats such as PNG and JPEG 2000 that are
 255 able to preserve data integrity as well as to provide
 efficient compression ratios for access and
 dissemination.

Metadata Representation

260 Metadata are commonly defined as “data about data.”
 They constitute the documentation or a structured
 description associated with a document. Image files
 automatically include a certain amount of metadata
 that are stored in an area of the file defined by the file
 265 format and called *the header* but information may also
 be stored externally.
 In the widely used TIFF format, the term “tagged”
 indicates that developers can define and apply
 dedicated tags to enable them to include their own
 270 proprietary information (called “private tags”) inside
 a TIFF file without causing problems of compatibility.
 More recently, Exif format (Exchangeable image file
 format) was created by the Japan Electronic Industries
 Development Association. The latest version was
 275 published in 2002 and while the specification is not
 currently maintained by any industry or standards
 organization, its use by camera manufacturers is nearly
 universal. Exif fields are generated at the creation of
 the image and should not be modified after with the
 280 aim of including additional information like title or

keywords. To do this, other formats are recommended,
 such as XMP (eXtensible Metadata Platform). This last
 is an XML-based standard for creating, processing and
 storing standardized, extensible and proprietary
 285 metadata, created by Adobe Systems in 2001. XMP
 metadata can be embedded into a significant number of
 popular file formats. It is used in PDF and other image
 formats such as JPEG, JPEG 2000, GIF, PNG, TIFF
 and EPS.

290 In parallel to generic standards, standards dedicated to
 specific image applications also exist, such as DICOM
 (Digital Imaging and Communications in Medicine).
 This is a standard created in 1992 and widely adopted
 by hospitals, for handling, storing, printing and
 295 transmitting information in medical imaging. It
 includes a file format definition and a network
 communications protocol.

Metadata constitute the documentation of all aspects of
 digital files essential to their persistence, usefulness
 300 and access. Images without appropriate metadata may
 become hard to view, migrate to new technology, or to
 access among large volumes of images. When
 annotation is inappropriate or is missing, the
 representation of the visual content of images by image
 305 analysis may be an interesting alternative.

Image Content Representation

Born in the early 1990s, Content-Based Image
 Indexing (CBIR) is a discipline that exploits
 310 techniques of image analysis and databases. Indexing
 an image by its content consists of automatically
 extracting structures that describe the visual content
 relevantly for the considered application. These
 structures can describe the visual content of an image
 315 globally or locally by characterizing its distribution of
 color, shape and texture, or parts or objects of the
 image. The visual structures exhibited are considered
 as the index of the image, they are digitally represented
 with one or several multidimensional vectors called
 320 *signature* of the image.

Key Applications

JPEG 2000 is a standard that gathers an image file
 325 format and an algorithm of image compression, created
 by the Joint Photographic Experts Group committee in
 2000. The coding algorithm of JPEG 2000 is similar to
 the JPEG one. It mainly differs in the use of wavelets
 instead of a DCT. Wavelets provide a decomposition
 330 of the image into a pyramid of sub-images which store
 different levels of resolution of the image. They can be
 of two types, according to the objective:
 (1) a Daubechies wavelet transform that requires
 quantization to reduce the amount of bits representing
 335 data, as JPEG does, and then imposes lossy
 compression;

(2) a rounded version of Le Gall wavelet transform, that uses only integer coefficients and then does not require quantization, providing lossless coding.

340 The aim of this standard is not only improving compression performance but also adding features, among which are transmission error resilience and region of interest (ROI). This last offers the opportunity of storing parts of the same picture using

345 different quality. Some parts of particular interest such as faces can be stored with higher quality, to the detriment of other ones where low quality/high compression can be tolerated. The JPEG 2000 standard defines two file formats that support embedded XML

350 metadata: JP2, which supports simple XML, and JPX, which has a more robust XML system based on an embedded metadata initiative of the International Imaging Industry Association (the DIG35 specification).

355

Future Directions

Today, there is a need of a unified framework for the creation, representation, storage, access, delivery,

360 management and protection of multimedia contents. New standards for managing these contents are the international standards MPEG-21 and MPEG-7. MPEG-21 is a standard started in 1999 by MPEG

365 (Moving Picture Experts Group) and now normalized as ISO/IEC 21000. Its main objectives are to define an open framework for multimedia applications and more precisely to provide a standardized structure for various media contents and to facilitate their access, delivery, management and protection. MPEG-7 is

370 another ISO/IEC standard started by MPEG in 1998 and formally called Multimedia Content Description Interface. One of its main objectives is to provide unified and efficient searching, filtering and content identification methods for these media.

375

(Abridged)

Recommended Reading

- 380 1. Adobe XMP main page:
<http://www.adobe.com/products/xmp/index.html>. 5
2. Datta R. Joshi D. Li J. and Wang J. Z. Image retrieval: Ideas, influences, and trends of the new age. ACM Comput. Surv. 40 (2), 5 April 2008.
- 385 3. File format info.
<http://www.fileformat.info/format/could.htm>. 3
4. Ian S., Burnett, Pereira F., Van de Walle R., and Koenen R. The MPEG-21 Book. Wiley, 462 pages, 6 May 2006.
- 390 5. Manjunath B.S., Salembier P., and Sikora T. Introduction to MPEG-7: Multimedia Content Description Interface. Wiley & Sons, 6 April 2002.

6. Murray J. D. and van Ryper W. Encyclopedia of Graphics File Formats. O'Reilly, 1152 pages, 2nd edition, 1996.
- 395 7. Oleg S., Pinykh. Digital Imaging and Communications in Medicine (DICOM): A Practical Introduction and Survival Guide. Springer, 2008.
8. Salomon D., Motta G., and Bryant D. Data
- 400 Compression: The Complete Reference. Springer, 4th edition, 2006.
9. Taubman D. S. and Marcellin M. W. JPEG 2000: Image Compression Fundamentals, Standards and Practice. Kluwer International Series in Engineering and Computer Science, Secs 642, 52001.
- 405 10. Wallace G. K. The JPEG still picture compression standard. Commun. of the ACM, 34(4):30–44, 3 April 1991.

Answer the following questions:

- 1) What are the two main categories of digital images and what is the difference between them?
- 2) What are the key factors in determining image quality?
- 3) What is compression and what types of compression are mentioned in the text?
- 4) Name some types of lossless compression.
- 5) What is JPEG and what is it used for?
- 6) Name some vector image formats.
- 7) What are file formats in which bitmap data can be saved?
- 8) What term is used for “data about data”?
- 9) Why do we need content-based image indexing?
- 10) In what way does JPEG 2000 differ from the JPEG standard?
- 11) What is the standard for multimedia applications?

Match the following terms with their definitions:

- 1) Run-length encoding (RLE)
 - 2) Bitmapped image
 - 3) Content-based information retrieval
 - 4) Color model
 - 5) Lossy compression
 - 6) TIFF
 - 7) JPEG 2000
-
- a) A file format that uses wavelet compression
 - b) An attempt to describe color in a mathematical, predictable and reproducible way
 - c) An image made up of a given number of pixels, each with a specific color value, laid out in a grid
 - d) Technology that is able to retrieve images on the basis of machine-recognizable visual criteria
 - e) Reduction in file size that involves permanent loss of information

- f) Large runs of consecutive identical data values replaced by a simple code with the data value and length of the run
- g) Tagged Image File Format

Mark the following statements as *true* or *false*:

- 1) Vector graphics are images that are completely described using mathematical definitions.
- 2) Vector drawings can usually be scaled without any loss in quality.
- 3) Common raster images include 2- and 3-D architectural drawings, flow charts, logos and fonts.
- 4) The Lempel-Ziv (LZ) compression methods are among the most popular algorithms for lossy storage.
- 5) JPEG 2000 is a newer version of the JPEG format that compresses images via wavelets.
- 6) PNG (Portable Network Graphics) was a predecessor of the GIF format for transfer and display images online.

Vocabulary

advent ['æd.vent], [-vənt] – příchod, nástup

band [bænd] – pásmo

column ['kɒl.əm] ^{US} ['kɑ:l.əm] – sloupec

consecutive [kən'sek.ju.tɪv] ^{US} [-tɪv] – následný

cosine ['kəʊ.saɪn] ^{US} ['kou-] – kosinus

curve [kɜ:v] ^{US} [kɜ:v] – křivka

dedicated ['ded.ɪ.keɪ.tɪd] ^{US} [-tɪd] – zaměřený, věnovaný

detriment ['det.rɪ.mənt] – škoda, újma

dimension [,daɪ'men.tʃən] – rozměr

dissemination [dɪ'sem.ɪ.neɪʃən] – šíření

edge [edʒ] – hrana

formula ['fɔ:m.ju.lə] ^{US} ['fɔ:r-] – vzorec, program, plán

framework ['freɪm.wɜ:k] ^{US} [-wɜ:k] – rámeček

geometrical primitives [,dʒi:.ə'met.rɪk.əl]

['pri.mi.tɪvz] – geometrické útvary

grid [grɪd] – mřížka

charge coupled device [tʃɑ:dʒ] ['kʌp.lɪ] [dɪ'vaɪs] ^{US}

[tʃɑ:rdʒ] [-] [-] – zařízení s vázanými náboji

intentionally [ɪn'ten.ʃən.əli] – záměrně

line [laɪn] – čára, přímka

lossless ['lɒslɪs] – bezztrátový

lossy ['lɒsi] – ztrátový

means [mi:nz] – prostředek

objective [əb'dʒek.tɪv] – cíl

occurrence [ə'kʌr.ənts] ^{US} [-'kɜ:-] – výskyt

perceptual [pə'sep.ʃu.əl] – vnímavostní

persistence [pə'sɪs.tənʃs] ^{US} [pə-] – stálost, vytrvalost

point [pɔɪnt] – bod

polygon ['pɒl.ɪ.gɒn] ^{US} ['pɑ:l.ɪ.gɑ:n] –

mnohoúhelník, polygon

probability [,prɒb.ə'bɪl.ɪ.ti] ^{US} [,pra: .bə'bɪl.ə.ti] – pravděpodobnost

proprietary [prə'praɪ.ə.tri] ^{US} [-ter.ɪ] – vlastnický, patentovaný

replacement for [rɪ'pleɪs.mənt] – náhrada za

representation [,rep.rɪ.zen'teɪ.ʃən] – zobrazení, vyobrazení

resilience [rɪ'zɪl.i.ənts] – odolnost

row [rəʊ] ^{US} [rou] – řádek

run-length coding [rʌn] [lenk θ] [kəʊdɪŋ] ^{US} [-] [-] [kəʊdɪŋ] – kódování délkou běhu

scalable ['skeɪ.lə.bl] – škálovatelný

scriptable ['skrip.tə.bl] – skriptovatelný

shade [ʃeɪd] – odstín

shape [ʃeɪp] – tvar, útvar

smoothly ['smu:ð.li] – hladce

strength [streŋθ] – síla, intenzita

string [strɪŋ] – řetězec

to allow for [ə'laʊ] – počítat s čím

to associate with [ə'səʊ.si.eɪt] ^{US} [-'səʊ-] – spojovat s

to be made up of [meɪd] – být složen

to capture ['kæp.tʃə] ^{US} [-tʃə] – zachytit

to convey [kən'veɪ] – sdělit, vyjádřit

to differ in st ['dɪf.ə] ^{US} [-ə] – lišit se v něčem

to discard [dɪ'ska:d] ^{US} [-'ska:rd] – odhodit

to distinguish [dɪ'stɪŋ.gwɪʃ] – rozlišit

to embed st in st [ɪm'bed] – zapustit, zasadit do

to employ [ɪm'plɔɪ] – použít, využít

to exploit [ɪk'splɔɪt] – využít, zužítkovat

to facilitate [fə'sɪl.ɪ.teɪt] – usnadnit

to focus on st ['fəʊ.kəs] ^{US} ['fou-] – zaměřit se na

to challenge ['tʃæl.ɪndʒ] – zpochybnit, vznést námitky

to migrate [maɪ'greɪt] – přesunout, přemístit

to mitigate ['mɪt.ɪ.geɪt] ^{US} ['mɪt-] – zmírnit

to neglect st [nɪ'glekt] – zanedbat, opomenout

to preserve [prɪ'zɜ:v] ^{US} [-'zɜ:v] – uchovat

to render ['ren.də] ^{US} [-də] – zobrazit

to retain [rɪ'teɪn] – uchovat

to round [raʊnd] – zaokrouhlit

to sacrifice ['sæk.rɪ.faɪs] – obětovat

to settle on st ['set.l] ^{US} ['set-] – rozhodnout se pro

to shrink [ʃrɪŋk] – zmenšit

to stand for [stænd] – znamenat

to tie [taɪ] – vázat se, souviset

to treat st as [tri:t] – považovat co za

tuple [tju:pəl] – n-tice

unlike [ʌn'laɪk] – na rozdíl

versatile ['vɜ:s.ə.taɪl] ^{US} ['vɜ:s.ə.tə] – všestranný, univerzální

Phrases

in a zigzag fashion ['zɪg.zæɡ] ['fæʃ.ən] – klikatě, cikcak

in contrast to ['kɒn.trɑːst] ⓘ ['kɑːn.træst] – naproti tomu, na rozdíl

in parallel to ['pær.ə.lel] ⓘ ['per-] – souběžně

in particular [pə'tɪk.jʊ.lər] ⓘ [pə'tɪk.jə.lər] – obzvláště

widely used – hojně využíváný

with the aim of agreeing – s cílem shodnout se

Processor Cache

PETER BONCZ
CWI, Amsterdam, The Netherlands

Definition

To hide the high latencies of DRAM access, modern computer architecture now features a memory hierarchy that besides DRAM also includes SRAM cache memories, typically located on the CPU chip. Memory access first checks these caches, which takes only a few cycles. Only if the needed data is not found, an expensive memory access is needed.

Key Points

CPU caches are SRAM memories located on the CPU chip, intended to hide the high latency of accessing off-chip DRAM memory. Caches are organized in cache lines (typically 64 bytes). In a fully-associative cache, each memory line can be stored in any location of the cache. To make checking the cache fast, however, CPU caches tend to have limited associativity, such that storage of a particular cache line is possible in only 2 or 4 locations. Thus only 2 or 4 locations need to be checked during lookup (these are called 2- resp. 4-way associative caches). The cache hit ratio is determined by the spatial and temporal locality of the memory access generated by the running program(s).

Cache misses can either be compulsory misses (getting the cache lines of all used memory once), capacity misses (caused by the cache being too small to keep all multiply used lines in cache), or conflict misses (due to the limited associativity of the cache).

Most modern CPUs have at least three independent caches: an instruction cache to speed up executable instruction fetch, a data cache to speed up data fetch and store, and a Translation Lookaside Buffer (TLB) used to speed up virtual-to-physical address translation for both executable instructions and data. The TLB is not organized in cache lines; it simply holds pairs of (virtual, logical) page mappings, typically a fairly limited amount (e.g., 64). In practice, this means that algorithms that repeatedly touch memory in more than 64 pages (whose size is often 4 KB) shortly after each other, run into TLB thrashing. This problem can sometimes be mitigated by setting a large virtual memory page size, or by using special large OS pages (sometimes supported in the CPU with a separate, smaller, TLB for large pages).

Another issue is the tradeoff between latency and hit rate. Larger caches have better hit rates but longer latency. To address this tradeoff, many computers use multiple levels of cache, with small fast caches backed up by larger slower caches. Multi-level caches

generally operate by checking the smallest Level 1 (L1) cache first; if it hits, the processor proceeds at high speed. If the smaller cache misses, the next larger cache (L2) is checked, and so on, before external memory is checked. As the latency difference between main memory and the fastest cache has become larger, some processors have begun to utilize as many as three levels of on-chip cache.

For multi-CPU and multi-core systems, the fact that some of the higher levels of cache are not shared, yet provide coherent access to shared memory, causes additional cache-coherency inter-core communication to invalid stale copies of cache lines on other cores when one core modifies it. In multi-core CPUs, an important issue is that cache level is shared among all cores – this cache level is on the one hand a potential hot-spot for cache conflicts, on the other hand provides an opportunity for very fast inter-core data exchange.

In case of sequential data processing, the memory controller or memory chipset in modern computers often detects this access pattern and starts requesting the subsequent cache lines in advance. This is called hardware prefetching. Prefetching effectively allows hiding compulsory cache misses. Without prefetching, the effective memory bandwidth would equate cache line size divided by memory latency (e.g., $64/50\text{ns} = 1.2\text{ GB/s}$). Thanks to hardware prefetching, modern computer architectures reach four times that on sequential access. Modern CPUs also offer explicit prefetching instructions, which a software writer can exploit to perform (non-sequential) memory access in advance, hiding their latency. In database systems, such software prefetching has successfully been used in making hash-table lookup faster (e.g., in hash-join and hash aggregation).

In database systems, a series of cache-conscious data storage layouts (e.g., DSM and PAX) have been proposed to improve cache line usage. Also, a number of cache-conscious query processing algorithms, such as cache-partitioned hash join and hash-join using memory prefetching, have been studied. In the area of data structures and theoretical computer science, there has recently been interest in cache-oblivious algorithms that regardless of the exact parameters of the memory hierarchy (number of levels, cache size, cache line sizes and latencies) perform well.

(Abridged)

Answer the following questions:

- 1) What kinds of independent caches are mentioned in the text?
- 2) What is the cache hit ratio determined by?
- 3) What kinds of cache misses are mentioned in the text?
- 4) What are cache-oblivious algorithms?
- 5) How do multi-level caches operate?
- 6) What enables a faster hash-table lookup?

Match the following terms with their definitions:

- 1) hardware prefetching
 - 2) cache miss
 - 3) fully-associative cache
 - 4) translation lookaside buffer
-
- a) this type of cache speeds up virtual-to-physical address translation
 - b) requesting the subsequent memory lines in advance
 - c) a situation when the requested data does not occur in the cache
 - d) in this type of cache each memory line can be stored in any location of the cache

Mark the following statements as *true* or *false*:

- 1) Modern computer architecture does not include DRAM, only SRAM.
- 2) A big problem with DRAM is its high latency.
- 3) A translation lookaside buffer is organized in cache lines.
- 4) Larger caches have worse hit rates but lower latency.

Vocabulary

Coherent [kəʊ 'hiə.rənt] ^{US} [kou'hir.ənt] – souvislý, soudržný
Compulsory [kəm'pʌl.səri] ^{US} [-sə-] – povinný
conscious ['kɒn.tʃəs] ^{US} ['ka:n-] – vědom si něčeho
fairly ['feə.li] ^{US} ['fer-] – celkem, dost
hotspot [hɒtspɒt] ^{US} [hɒ:tspɒ:t] – ohnisko
independent [ɪn.dɪ'pen.dənt] – nezávislý
invalid [ɪn'væl.ɪd] – neplatný
latency ['leɪ.tən.t.si] – zpoždění
oblivious [ə'blɪv.i.əs] – zapomnětlivý, zapomínající, nedbající (něčeho)
ratio ['reɪ.fɪ.əʊ] ^{US} [-oʊ] – poměr
sequential [sɪ'kwɛn.fəl] – následující, postupný
spatial ['speɪ.ʃəl] – prostorový
stale [steɪl] – *zde* zastaralý
subsequent ['sʌb.sɪ.kwənt] – následující, příští
table ['teɪ.bl] – tabulka
temporal ['tem.pər.əl] ^{US} [-pə.əl] – časový
thrashing [θræʃɪŋ] – zahlcení paměti
to equate [ɪ'kweɪt] – rovnat se
to fetch [fetʃ] – přinést
to mitigate ['mɪt.i.geɪt] ^{US} ['mɪtʃ-] – zmírnit
to multiply ['mʌl.tɪ.plaɪ] ^{US} [-tʃɪ-] – násobit
to propose [prə'pəʊz] ^{US} [-'pouz] – navrhnout
to tend to [tend] – mít tendenci k
to utilize ['ju:.tɪ.laɪz] ^{US} [-tʃəl.aɪ-] – využít, upotřebit
trade-off ['treɪd.ɒf] ^{US} [-ɑ:f] – kompromis

Phrases

in advance [əd'vɑ:nts] ^{US} [-'vænts] – dopředu, předběžně, v předstihu
regardless of [rɪ'gɑ:d.ləs] ^{US} [-'gɑ:rd-] – bez ohledu na

Spatial Data Mining

SHASHI SHEKHAR, JAMES KANG,
VIJAY GANDHI

5 University of Minnesota, Minneapolis, MN, USA

Definition

10 Spatial data mining is the process of discovering nontrivial, interesting, and useful patterns in large spatial datasets. The most common spatial pattern families are co-locations, spatial hotspots, spatial outliers, and location predictions.

15 Historical Background

Spatial data mining research began several decades ago when practitioners and researchers noticed that critical assumptions in classical data mining and statistics were violated by spatial datasets. First, whereas classical datasets often assume that data are discrete, spatial data were observed to reside in continuous space. For example, classical data mining and statistical methods may use market-basket datasets (e.g., history of Walmart's transactions), where each item-type in a transaction is discrete. However, "transactions" are not natural in continuous spatial datasets, and decomposing space across transactions leads to loss of information about neighbor relationships between items across transaction boundaries. In addition, spatial data often exhibits heterogeneity (i.e., no places on the Earth are identical), whereas classical data mining techniques often focus on spatially stationary global patterns (i.e., ignoring spatial variations across locations). Finally, one of the common assumptions in classical statistical analysis is that data samples are independently generated. However, this assumption is generally false when analyzing spatial data, because spatial data tends to be highly self-correlated. For example, people with similar characteristics, occupation and background tend to cluster together in the same neighborhoods. In spatial statistics this tendency is called spatial auto-correlation. Ignoring spatial auto-correlation when analyzing data with spatial characteristics may produce hypotheses or models that are inaccurate or inconsistent with the data set. Thus, classical data mining algorithms often perform poorly when applied to spatial data sets. Better methods are needed to analyze spatial data to detect spatial patterns.

Foundations

55 The spatial data mining literature has focused on four main types of spatial patterns: (i) spatial outliers, which are spatial locations showing a significant

60 difference from their neighbors; (ii) spatial co-locations, or subsets of event types that tend to be found more often together throughout space than other subsets of event types; (iii) location predictions, that is, information that is inferred about locations favored by an event type based on other explanatory spatial variables; and (iv) spatial hotspots, unusual spatial groupings of events. The remainder of this section presents a general overview of each of these pattern categories.

70 A *spatial outlier* is a spatially referenced object whose non-spatial attribute values differ significantly from those of other spatially referenced objects in its spatial neighborhood. For example, consider spatial outliers, detected in traffic measurements for sensors on highway I-35W (North bound) in the Minneapolis-St. Paul area, for a 24-h time period. Station 9 may be considered a spatial outlier as it exhibits inconsistent traffic flow compared with its neighboring stations. Once a spatial outlier is identified, one may proceed with diagnosis. For example, the sensor at Station 9 may be diagnosed as malfunctioning. Spatial attributes are used to characterize location, neighborhood, and distance. Non-spatial attribute dimensions are used to compare a spatially referenced object to its neighbors. Spatial statistics literature provides two kinds of bipartite multidimensional tests, namely graphical tests and quantitative tests. Graphical tests, such as Variogram clouds and Moran scatterplots, are based on the visualization of spatial data and highlight spatial outliers. Quantitative methods provide a precise test to distinguish spatial outliers from the remainder of data.

90 *Spatial co-location* pattern discovery finds frequently co-located subsets of spatial event types given a map of their locations. Spatial co-location is a generalization of a classical data mining pattern family called association rules, since transactions are not natural in spatial datasets, and partitioning space across transactions leads to loss of information about neighbor relationships between items near transaction boundaries. Additional details about co-location interest measures, e.g. participation index and K functions, and mining algorithms are described in [2].

100 *Location prediction* is concerned with the discovery of a model to infer preferred locations of a spatial phenomenon from the maps of other explanatory spatial features. For example, ecologists may build models to predict habitats for endangered species using maps of vegetation, water bodies, climate, and other related species. For example, consider an example of a dataset used in building a location prediction model for red-winged blackbirds in the Darr and Stubble wetlands on the shores of Lake Erie in Ohio, USA. This dataset consists of nest location, distance to open

water, vegetation durability and water depth maps. Classical prediction methods may be ineffective in this problem due to the presence of spatial auto-correlation. Spatial data mining techniques that capture the spatial autocorrelation of nest location such as the Spatial Autoregressive Model (SAR) [1] and Markov Random Fields based Bayesian Classifiers (MRF-BC) are used for location prediction modeling.

Spatial Hotspots are unusual spatial groupings of events that tend to be much more closely related than other events. Examples of spatial hotspots can be incidents of crime in a city or outbreaks of a disease. Hotspot patterns have properties of clustering as well as anomalies from classical data mining. However, hotspot discovery [4] remains a challenging area of research due to variation in shape, size, density of hotspots and underlying space (e.g., Euclidean or spatial networks such as roadmaps). Additional challenges arise from the spatio-temporal semantics such as emerging hotspots, displacement etc.

135 **Key Applications**

Spatial data mining and the discovery of spatial patterns has applications in a number of areas. Detecting spatial outliers is useful in many applications of geographic information systems and spatial databases, including the domains of public safety, public health, climatology, and location-based services. As noted earlier, for example, spatial outlier applications may be used to identify defective or out of the ordinary (i.e., unusually behaving) sensors in a transportation system. Spatial co-location discovery is useful in ecology in the analysis of animal and plant habitats to identify co-locations of predator-prey species, symbiotic species, or fire events with fuel and ignition sources. Location prediction may provide applications toward predicting the climatic effects of El Nino on locations around the world. Finally, identification of spatial hotspots can be used in crime prevention and reduction, as well as in epidemiological tracking of disease.

(Abridged)

160 **Recommended Reading**

1. Cressie N.A. *Statistics for Spatial Data* (Revised Edition). Wiley, New York, NY, 1993.
2. Huang Y., Shekhar S., and Xiong H. Discovering co-location patterns from spatial datasets: a general approach. *IEEE Trans. Knowl. Data Eng. (TKDE)*, 16(12):1472–1485, 2004.
3. Shekhar S., Schrater P., Vatsavai R., Wu W., and Chawla S. Spatial contextual classification and

- 170 prediction models for mining geospatial data. *IEEE Trans. Multimed. (special issue on Multimedia Databases)*, 4(2):174–188, 2002.
4. US Department of Justice - Mapping and Analysis for Public Safety report. *Mapping Crime: Understanding Hot Spots*, 2005
- 175 (<http://www.ncjrs.gov/pdffiles1/nij/209393.pdf>).
5. Shekhar S. and Chawla S. *A Tour of Spatial Databases*. Prentice Hall, 2003.
6. Longley P.A., Goodchild M., Maquire D.J., and Rhind D.W. *Geographic Information Systems and Science*. Wiley, 2005.
- 180 7. Shekhar S., Zhang P., Huang Y., and Vatsavai R. Trend in spatial data mining. In *Data Mining: Next Generation Challenges and Future Directions*, H. Kargupta, A. Joshi, K. Sivakumar, and Y. Yesha (eds.).
- 185 AAAI/MIT Press, 2003.
8. Solberg A.H., Taxt T., and Jain A.K. A Markov random field model for classification of multisource satellite imagery. *IEEE Trans. Geosci. Remote Sens.*, 34(1):100–113, 1996.
- 190 9. Kou Y., Lu C.T., and Chen D. Algorithms for spatial outlier detection. In *Proc. 2003 IEEE Int. Conf. on Data Mining, 2003*, pp. 597–600.
10. Shekhar S., Lu C.T., and Zhang P. A unified approach to detecting spatial outliers. *GeoInformatica*, 7(2):139–166, 2003.
- 195 11. Mamoulis N., Cao H., and Cheung D.W. Mining frequent spatio-temporal sequential patterns. In *Proc. 2003 IEEE Int. Conf. on Data Mining, 2005*, pp. 82–89.

Answer the following questions:

- 1) What is spatial data mining?
- 2) What are the basic differences between classical datasets and spatial datasets?
- 3) Why are classical data mining algorithms unsuitable for spatial datasets?
- 4) What does it mean that spatial data tends to be highly self-correlated? How would you explain spatial auto correlation?
- 5) Name the four main types of spatial patterns.
- 6) Explain a spatial outlier and give a simple example.
- 7) What are spatial and non-spatial attributes?
- 8) Which pairs are frequently co-located?
- 9) Give some examples of hotspots.
- 10) Name some areas where spatial data mining is applied.

Mark the following statements as *true* or *false*:

- 1) Many of the relationships on spatial data are implicit.
- 2) Spatial autocorrelation means that nearby things are more different than distant things.
- 3) In spatial data mining data samples are not independent.
- 4) Classical data mining techniques perform well when applied to spatial data sets.
- 5) Spatial data mining is based on the same assumptions as classical data mining.

Match the following terms with their definitions:

- 1) Spatial data mining (SDM)
 - 2) Spatial outlier
 - 3) Location prediction
 - 4) Spatial co-location
 - 5) Hotspot analysis
 - 6) Spatial autocorrelation
 - 7) Variogram clouds and Moran scatterplots
-
- a) Presence of two or more spatial objects at the same location or at significantly close distances from each other
 - b) Process of finding unusually dense event clusters across time and space
 - c) Process of discovering interesting, useful, non-trivial patterns from large spatial datasets
 - d) Spatial outlier detection algorithms based on visualization
 - e) Observation which appears to be inconsistent with its neighborhood
 - f) Prediction of events occurring at particular geographic locations
 - g) Spatial data values are influenced by values in their immediate vicinity

Vocabulary

anomaly [ə'nɒm.ə.li] ⑤ [-'nɑ: mə-] – odchylka
assumption [ə'sʌmp.ʃən] – předpoklad, domněnka
boundary ['baʊn.dər.i] [-dri] ⑤ [-dæ-] – hranice, mez
defective [di'fek.tɪv] – vadný, chybný
discrete [di'skri:t] – nespojitý, oddělený, samostatný
displacement [di'spleɪs.mənt] – přemístění, posun
endangered [ɪn'deɪn.dʒəd] ⑤ [-dʒæd] – ohrožený
grouping ['gru:pɪŋ] – seskupení
habitat ['hæb.ɪ.tæt] – přirozené prostředí, domov
hypothesis (pl. hypotheses) [haɪ'pɒθ.ə.sɪs] ⑤ [-'pɑ:θə-], pl. [haɪ'pɒθ.ə.sɪ:z] – hypotéza, předpoklad, domněnka
ignition [ɪg'nɪʃ.ən] – vznícení, vzplanutí
incident ['ɪnɪ.sɪ.dənt] – případ, incident, příhoda, událost
inconsistent [ɪn.kən'sɪs.tənt] – neslučitelný, jsoucí v rozporu
namely ['neɪm.li] – a to, jmenovitě
nontrivial [nɒn'trɪv.i.əl] – podstatný, důležitý
out of the ordinary [aʊt] ['ɔ:di.nə.ri] ⑤ [-] ['ɔ:r.dən.ər-] – neobvyklý, zvláštní
outbreak ['aʊt.breɪk] – vypuknutí
outlier ['aʊt.laɪ.ə] – co leží mimo
pattern ['pæt.ən] ⑤ ['pæʃ.ən] – vzorec, způsob, průběh
phenomenon (pl. phenomena) [fə'nɒm.i.nən] ⑤ [-'nɑ: mə.nɑ:z], pl. [fə'nɒm.i.nə] – jev
predator ['pred.ə.tər] ⑤ [-tə] – dravec, predátor
prey [preɪ] – kořist
remainder [rɪ'meɪn.dər] ⑤ [-dər] – zbytek
species ['spi:ʃi:z] – druh
stationary ['steɪ.ʃən.ər.i] ⑤ [-ʃə.nər-] – neměnný, nehybný
throughout space [θru:'aʊt][speɪs] – po celém prostoru
to be concerned with st [kən'sɜ:nd] ⑤ [-'sɜ:nd] – zabývat se čím
to cluster ['klʌs.tər] ⑤ [-tər] – shlukovat se, seskupit
to co-locate [kəʊ.ləʊ'keɪt] ⑤ [kəʊ.ləʊ'keɪt] – vyskytovat se společně

to decompose [di:kəm'pəʊz] ⑤ [-'pəʊz] – rozložit

to favor ['feɪ.vər] ⑤ [-vər] – podporovat

to highlight ['haɪ.laɪt] – zvýraznit, poukázat

to infer [ɪn'fɜ:z] ⑤ [-'fɜ:z] – dedukovat, vyvozovat

to note [nəʊt] ⑤ [nəʊt] – upozornit, poukázat

to partition [pɑ:'tɪʃ.ən] ⑤ [pɑ:r-] – rozdělit

to proceed [prəʊ'si:d] ⑤ [prəʊ-] – pokračovat, postupovat

to reference ['ref.ər.ənʃs] ⑤ [-ə-] – odkazovat

to tend [tend] – mít sklon, tendenci, být náchylný

to track [træk] – sledovat

to violate ['vaɪə.leɪt] – porušit, nedodržet

Phrases

due to – kvůli

either – or – buď – nebo

from the standpoint ['stænd.pɔɪnt] – z hlediska

in infant stages ['ɪn.fənt] – na počátku vývoje

Stemming

CHRIS D. PAICE
Lancaster University, Lancaster, UK

Definition

Stemming is a process by which word endings or other affixes are removed or modified in order that word forms which differ in non-relevant ways may be merged and treated as equivalent. A computer program which performs such a transformation is referred to as a stemmer or stemming algorithm. The output of a stemming algorithm is known as a stem.

Historical Background

The need for stemming first arose in the field of information retrieval (IR), where queries containing search terms need to be matched against document surrogates containing index terms. With the development of computer-based systems for IR, the problem immediately arose that a small difference in form between a search term and an index term could result in a failure to retrieve some relevant documents. Thus, if a query used the term “explosion” and a document was indexed by the term “explosives,” there would be no match on this term (whether or not the document would actually be retrieved would depend on the logic and remaining terms of the query). The first stemmer for the English language to be fully described in the literature was developed in the late 1960s by Julie Beth Lovins [11]. This has now been largely superseded by the Porter stemmer [14], which is probably the most widely used, and the Paice/Husk stemmer [12]. Stemmers have also been developed for a wide variety of other languages.

Foundations

Definitions

In an IR context, the process of taking two distinct words, phrases or other expressions and treating them as semantically equivalent is referred to as conflation. The two expressions need not be precisely synonymous, but they must refer to the same core concept (compare “computed” and “computable”). In this article, the term “practically equivalent” is used to mean that, for the purposes of a particular application, the words may as well be taken as equivalent. The term conflation is sometimes used as though it is equivalent to stemming, but it is in fact a much broader concept, since it includes (i) cases where the strings concerned are multi-word expressions, as in “access time” and “times for access”, and (ii) cases where the strings are not etymologically related, as in

“index term” and “descriptor”. In case (i) special string matching techniques may be used, whereas in case (ii) reference to a dictionary or thesaurus is necessary. The present account deals exclusively with the conflation of single etymologically related words. There are various possible approaches to word conflation, including the following.

1. Direct matching. In this method, the character sequences of two words are compared directly, and a similarity value is computed. The words are then considered to match if their mutual similarity exceeds a predefined threshold. To give a simple example, the first six letters of the words “exceeds” and “exceeded” are the same, so these words together contain 12 matching letters out of 15. Hence, a similarity of $12/15 = 0.80$ can be computed. Use of a threshold (say, 0.70) allows a decision as to whether the words can be considered equivalent. With such a method, setting the threshold is problematic. Thus, the similarity between “exceeds” and “excess” is 0.62, which is below the stated threshold. However, allowing for this by lowering the threshold to 0.60 would cause “excess” and “except” (similarity 0.67) to be wrongly conflated.

2. Lexical conflation. In this case a thesaurus or dictionary is used to decide whether two words are equivalent. Obviously, this method can be used even for etymologically unrelated words. A problem here is obtaining a suitably comprehensive and up-to-date thesaurus, and one which explicitly lists routine variants such as plurals.

3. Cluster-based conflation. This method, investigated by Xu and Croft [15], involves creating clusters of practically equivalent words by analyzing the word associations in a large representative text corpus. Each query word is then supplemented by adding in the other words in its cluster. In contrast to method (2), the clusters created are specific to the text collection in question. However, the creation of the clusters can be very time-consuming.

4. N-gram conflation. In this method, each word is decomposed into a collection of N-letter fragments (N-grams), and a similarity is computed between the N-gram collections of two words; a threshold is then applied to decide whether the words are equivalent. This approach was pioneered by Adamson and Boreham[1], who used sets of bigrams, where $N = 2$. For example, after eliminating duplicates and sorting into order, “exceeds” can be represented by the bigram set {ce, ds, ed, ee, ex, xc} and “exceeded” by {ce, de, ed, ee, ex, xc}. Out of 7 distinct bigrams here,

5 are shared between the two words; hence a similarity of $5/7 = 0.712$ can be computed.

115 5. Stemming. Stemming refers to the removal of any
suffixes (and sometimes other affixes) from an input
word to produce a stem. Two words are then deemed
to be equivalent if their stems are identical. This
method is much favored because it is fast: all words
120 can be reduced to stems on input to the system, and
simple string matching used thereafter. The remainder
of this article focuses on stemming in this narrow
sense.

125 Prefixes and Infixes

In English, stemmers are usually designed for
removing suffixes from words. The removal of
“intimate” prefixes such as “intro-,” “pro-” and
130 “con-” generally results in words being wrongly
conflated (consider “intro-duction,” “pro-duction”
and “con-duction”).

However, there may be a case for removing looser
prefixes such as “hyper-” or “macro-.” Also, prefix
135 removal may be desirable in certain domains with
highly artificial vocabularies, such as chemistry and
medicine. As explained below, there are some
languages in which removal or replacement of
prefixes, or even infixes, is in fact essential.

140

Performance and Evaluation

Since stemmers were originally developed to aid the
operation of information retrieval systems, it was
145 natural that they were first assessed in terms of their
effect on retrieval performance, as well as on
“dictionary compression” rates. Researchers were
frustrated to find that the effects on retrieval
performance for English language material were small
and often negative [10]. Removal of “-s” and other
150 regular inflectional endings might be modestly helpful,
but use of heavier stemming could easily result in a
loss of performance [7].

Stemming errors are of two kinds: understemming, in
155 which a pair of practically equivalent words are not
conflated, and overstemming, in which two
semantically distinct words are wrongly conflated.

Non-English Stemmers

160

Stemming is appropriate for most (though not all)
natural languages, and appears to be especially
beneficial for highly inflected languages [9]. There is
neither space nor need to describe non-English
165 stemmers here, except to note that some languages
exhibit much greater structural complexity, and this
warrants special approaches. Thus, a typical Arabic

word consists of a root verb of three (or occasionally
four or five) consonants (e.g., ‘k-t-b’ for ‘to write’),
170 into which various prefixes, infixes and suffixes are
inserted to produce specific variant forms (‘katabna’:
‘we wrote’ and ‘kitab’: ‘book’).

Some researchers have concentrated on extracting the
correct root from a word [3], but Aljlayl and Frieder
175 have demonstrated that better retrieval performance is
obtained by using a simpler “light stemming”
approach, in which only the most frequent suffixes and
prefixes are removed [4]. Their results showed that
extraction of roots causes unacceptable levels of
180 overstemming.

Key Applications

As noted earlier, stemmers are routinely used in
185 information retrieval systems to control vocabulary
variability. They also find use in a variety of other
natural language tasks, especially when it is required to
aggregate mentions of a concept within a document or
set of documents. For example, stemmers may be used
190 in constructing lexical chains within a text. Stemming
can also have a role to play in the standardization of
data for input to a data warehouse.

(Abridged)

195

Recommended Reading

1. Adamson G.W. and Boreham J. The use of an
association measure based on character structure to
200 identify semantically related pairs of words and
document titles. *Inf. Process. Manage.*, 10(7/8):253–
260, 1974.
2. Ahmad F., Yusoff M., and Sembok M.T.
Experiments with a stemming algorithm for Malay
words. *J. Am. Soc. Inf. Sci. Technol.*, 47(12):909–918,
205 1996.
3. Al-Sughaiyer I.A. and Al-Kharashi I.A. Arabic
morphological analysis techniques: a comprehensive
survey. *J. Am. Soc. Inf. Sci. Technol.*, 55(3):189–213,
210 2004.
4. Aljlayl M. and Frieder O. On Arabic search:
Improving the retrieval effectiveness via a light
stemming approach. In , 2002, pp. 340–347.
5. Bacchin M., Ferro N., and Melluci M. A
probabilistic model for stemmer generation. *Inf.*
215 *Process. Manage.*, 41(1):121–137, 2005.
6. Frakes W.B. and Fox C.J. Strength and similarity of
affix removal stemming algorithms. *SIGIR Forum*,
37(1):26–30, 2003 (Spring 2003).
- 220 7. Harman D. How effective is suffixing? *J. Am. Soc.*
Inf. Sci., 42(1):7–15, 1991.

8. Hull D. A Stemming algorithms: a case study for detailed evaluation. *J. Am. Soc. Inf. Sci.*, 47(1):70–84, 1996.
- 225 9. Krovetz R. Viewing morphology as an inference process. *Artificial Intelligence*, 118(1/2):277–294, 2000.
10. Lennon M., Pierce D.S., Tarry B.D., and Willett P. An evaluation of some conflation algorithms for information retrieval. *J. Inf. Sci.*, 3:177–183, 1981.
- 230 11. Lovins J.B. Development of a stemming algorithm. *Mech. Transl. Comput. Linguist.*, 11:22–31, 1968.
12. Paice C.D. Another stemmer. *SIGIR Forum*, 24(3):56–61, 1990.
- 235 13. Paice C.D. A method for the evaluation of stemming algorithms based on error counting. *J. Am. Soc. Inf. Sci.*, 47(8):632–649, 1996.
14. Porter M.F. An algorithm for suffix stripping. *Program*, 14(3):130–137, 1980.
- 240 15. Xu J. and Croft W.B. Corpus-based stemming using co-occurrence of word variants. *ACM Trans. Inf. Syst.*, 16(1):61–81, 1998.

Answer the following questions:

- 1) How would you describe *stemming*? What is its purpose?
- 2) What often resulted in a failure to retrieve relevant documents during searches in the past?
- 3) What is *conflation*?
- 4) Is there any difference between *conflation* and *stemming*?
- 5) What tools have to be used when strings are not etymologically related?
- 6) Describe *direct matching*.
- 7) What does the term of *threshold* refer to in the text?
- 8) What is the disadvantage of *cluster-based conflation*?
- 9) What are *bigrams*?

- 4) *Hyper-* in *hyperactive* is a suffix.
- 5) The term *affix* covers both *prefix* and *suffix*.
- 6) Stemming appears beneficial for highly inflected languages.
- 7) The *light-stemming approach* is based on removing the least frequent affixes.

Match the following terms and their definitions:

- 1) lexical conflation
 - 2) cluster-based conflation
 - 3) N-gram conflation
 - 4) understemming
 - 5) overstemming
-
- a) a method using a corpus of texts
 - b) a method based on bigrams
 - c) a situation where more-or-less equivalent words are not conflated
 - d) a method using a dictionary or thesaurus
 - e) a situation where two semantically distinct words are wrongly conflated

Mark the following statements as *true* or *false*:

- 1) During the conflation, the expressions need to be synonymous.
- 2) The words *mother* and *father* are etymologically related.
- 3) In stemming, two words are considered equivalent provided their stems are identical.

Vocabulary

account [ə'kaunt] – výčet; účet
actual ['æktʃu.əl] [-tju-] [-tʃʊl] – vlastní
actually ['æktʃu.ə.li] [-tju-] [-tʃʊ.li] – vlastně
affix ['æf.ɪks] – affix (předpona, přípona)
algorithm ['æl.gə.rɪ.ðəm] – algoritmus
bigram ['baɪgræm] – bigram (skupina dvou písmen, slabik či slov)
cluster ['klʌs.tə] ⓘ [-tə] – hrozen, skupina, klastř
comprehensive [kəm.pri'henʃ.sɪv] ⓘ [ka:m-] – komplexní, obsáhlý
conflation [kən'fleɪʃən] – spojování
compression [kəm'preʃ.ən] **rate** [reɪt] – kompresivita
consonant ['kɒn.sə.nənt] ⓘ ['ka:n-] – souhláska
core [kɔ:r] ⓘ [kɔ:r] – jádro; jádřinec
corpus ['kɔ:.pəs] ⓘ ['kɔ:r-] – korpus, tělo; soubor textů
distinct [dɪ'stɪŋkt] – různý, rozdílný
duplicate ['dju:.plɪ.kət] ⓘ ['du:-] – duplikát; duplikovaný (*srovnvej výslovnost s „to duplicate“* ['dju:.plɪ.keɪt] ⓘ ['du:-])
inflectional [ɪn'flekʃənəl] – skloňovací, skloňující, skloňovatelný
equivalent [ɪ'kwɪv.əl.ənt] – ekvivalentní
etymological [et.ɪ'mɒl.ə.dʒɪkəl] ⓘ [-'ma:lə-] – etymologický, vztahující se k původu slova
exclusive [ɪk'sklu:zɪv] – výhradní
failure ['feɪ.ljə] ⓘ [-ljə] – neúspěch
hence [henʃs] – tudíž
however [haʊ'ev.ə] ⓘ [-ə] – však, avšak
identical [aɪ'den.tɪ.kəl] ⓘ [-tə-] – identický, stejný
lexical ['lek.sɪ.kəl] – lexikální
lexical ['lek.sɪ.kəl] **chains** [tʃeɪn] – lexikální řetězce
loose [lu:z] – volný
mutual ['mju:.tʃu.əl] – vzájemný
predefined [pri:di'faɪnd] ⓘ [pri:də'faɪnd] – předem definovaný

prefix ['pri:fxɪks] – předpona
query ['kwɪə.ri] ⓘ ['kwɪr.i] – dotaz
remainder [rɪ'meɪn.də] ⓘ [-də] – zbytek
root [ru:t] – kořen
routine [ru:'ti:n] – obvyklý
semantic [sɪ'mæn.tɪk] ⓘ [-tɪk] – sémantický, významový
stem [stem] – kmen; stopka
surrogate ['sʌr.ə.gət] ⓘ ['sɜ:-] – náhradník, náhradní
suffix ['sʌf.ɪks] – přípona
synonymous [sɪ'nɒn.ɪ.məs] ⓘ [-'na:nə-] – podobného významu
thereafter [ðeə'ra:f.tə] ⓘ [ðer'æf.tə] – poté
thesaurus [θə'so:rəs] – thesaurus
threshold ['θreʃ.həʊld] ⓘ [-hould] – práh
thus [ðʌs] – tak, a tak
to aggregate st ['æg.rɪ.geɪt] – (na)hromadit něco
to aid st [eɪd] – napomáhat něčemu
to arise (arise, arose, arisen) [ə'raɪz] – objevit se; vyvstat
to assess st [ə'ses] – hodnotit něco
to conflate [kən'fleɪt] – spojit, spojovat
to decompose st [di:kəm'pəʊz] ⓘ [-'pouz] – rozložit něco
to deem [di:m] – považovat
to duplicate st ['dju:.plɪ.keɪt] ⓘ ['du:-] – duplikovat něco
to eliminate st [ɪ'lɪm.ɪ.neɪt] – eliminovat něco
to exceed st [ɪk'si:d] – překročit něco
to exhibit st [ɪg'zɪb.ɪt] – vykazovat něco
to extract st [ɪk'strækt] – extrahovat, vytáhnout něco
to favor st ['feɪ.və] ⓘ [-və] – dávat něčemu přednost
to investigate st [ɪn'ves.tɪ.geɪt] – vyšetřovat něco
to merge st [mɜ:dʒ] ⓘ [mɜ:dʒ] – spojit něco, sloučit
to obtain st [əb'teɪn] – získat něco
to pioneer [ˌpaɪə'niə] ⓘ [-'nɪr] – razit cestu
to retrieve st [rɪ'tri:v] – vyhledat, vyzvednout něco

to supersede st [ˌsuːpə'siːd] ⓘ [-pə-] - nahradit něco

to supplement st ['sʌp.lɪ.mənt] - doplnit něco

to treat st [tri:t] - zacházet s něčím

to warrant st ['wɒr.ənt] ⓘ ['wɔːr-] - opravňovat něco

variability [ˌveə.ri.ə'bɪl.ɪ.ti] ⓘ [ˌver.i.ə'bɪl.ə.ti] - variabilita

variant ['veə.ri.ənt] ⓘ ['ver.i-] - varianta

warehouse ['weə.haʊs] ⓘ ['wer-] - skladiště

whereas [weə'ræz] ⓘ [wer'æz] - kdežto

Phrases

In contrast to st ['kɒn.trɑːst] ⓘ ['kɑːn.træst] -
Oproti něčemu

Obviously, ... ['ɒb.vi.ə.sli] ⓘ ['ɑːb-] - Samozřejmě

Storage Devices

KALADHAR VORUGANTI

Network Appliance, Sunnyvale, CA, USA

Definition

One of the goals of database, file and block storage systems is to store data persistently. There are many different types of persistent storage devices technologies such as disks, tapes, DVDs, and Flash. The focus of this write-up is on the design trade-offs, from a usability standpoint, between these different types of persistent storage devices and not on the component details of these different technologies.

Historical Background

From a historical standpoint, tapes were the first type of persistent storage followed by disks, CDs, DVDs, and Flash. Newer types of memory technologies such as PRAM and MRAM are still in their infant stages. These newer non-volatile memory technologies promise DRAM access speeds and packaging densities, but these technologies are still too expensive with respect to cost/gigabyte.

Foundations

- Tapes/Tape Libraries: Tape readers/tape head, tape library, tape robot, and tape cartridge are the key components of a tape subsystem. Tapes provide the best storage packaging density in comparison to other types of persistent storage devices. Tapes do not provide random access to storage. Data on tapes can be stored either in compressed or uncompressed format. Unlike disks, tape cartridges can be easily transported between sites. Most organizations typically migrate data from older tape cartridges to newer tape cartridges every 5 years to prevent data loss due to material degradation. One can employ disk based caches in front of tape subsystems in order to allow for tapes to handle bursty traffic. Tapes that provide Write-Once, Read Many (WORM) characteristics are also available. WORM tapes are useful in data compliance environments where regulations warrant guarantees that a piece of data has not been altered. DLT and LTO are currently the two dominant tape technologies in the market. Technology-wise both these standards have minor differences. Finally, from a pure media cost standpoint, tapes are less expensive (cost per gigabyte) than disks and other forms of persistent media.

- Disks/Storage Controllers/NAS Boxes: Disks are the most widely used form of persistent storage media. Disks are typically accessed by enterprise level

applications when they are packaged as part of the processing server box (direct attached storage model), or are part of a network attached storage box (NAS) and accessed via NAS protocols or, are packaged as part of a storage controller box and accessed via storage area network protocols (SAN). The current trend is for protocol consolidation, where the same storage controller provides support for both SAN and NAS protocols. Typically, the size of the storage controllers can vary from a few terabytes to hundreds of terabytes (refrigerator sized storage controllers). A storage controller typically consists of redundant processors, protocol processing network cards, and RAID processing adapter cards. The disks are connected to each other via either arbitrated loop or switched networks. Storage controllers also contain multi-gigabyte volatile caches. Disks are also packaged as part of laptops.

There is a marked difference in the manufacturing process, and testing process between the enterprise class disks and commodity laptop class disks. Disks vary in their form factor, rotational speed, storage capacity, number of available ports, and the protocols used to access them. Currently, serial SCSI, parallel SCSI, serial ATA and parallel ATA, Fiber Channel, and SSA are the different protocols in use for accessing disks. Lower RPM and disk idle mode are new disk spin-down modes that allow disks to consume less power when they are not actively being used.

- DVD/Juke Boxes: DVDs and CDs are optical storage media that provide random access and WORM capabilities. Only recently, the multiple erase capacity of an individual CD or DVD was less than the capacity of a single disk drive or tape cartridge. DVDs can store more data than a CD, and a high definition DVD can store more data than a DVD. There are numerous competing standards for CDs, DVDs and high definition DVDs, however, format agnostic DVD players and DVD writers are emerging. Usage of DVDs is more prominent in the consumer space rather than in the enterprise space. A juke box system allows one to access a library of CDs or DVDs. DVDs have slower access speeds than most types of disks.

- Flash/SSDs/Hybrid Disks: Flash is memory technology that has non-volatile characteristics. Flash memory has slower read times than DRAM. Moreover, it has much slower write times than DRAM.

One has to perform an erase operation before one can re-use a flash memory location. One can only perform a limited number of erase operations. Thus, the number of write operations determines the Flash memory life. SLC and MLC are the two different NAND flash technologies. SLC can be erased a greater number of times, and it has faster access times than MLC based

flash. NAND flash has faster write and erase times than NOR flash. NOR flash has faster read times than NAND flash. NAND flash is used to store large amounts of data, whereas NOR flash is used to store executable code.

115
120
125
130
People are using MLC flash in cameras and digital gadgets, and are using SLC flash as part of solid state disks (SSDs). SSDs provide block level access interface (SCSI), and they contain a controller that performs flash wear leveling and block allocation. Hybrid disks that contain a combination of disks and Flash are emerging. Hybrid disks provide a Flash cache in front of the disk media. One typically can store meta-data or recently used data in the flash portion of hybrid disks to save on power consumption. That is, one does not have to spin up the disk. Flash storage provides much better random access speeds than disk based storage.

Key Applications

135
140
145
Tapes are being used primarily for archival purposes because they provide good sequential read/write times. Disks are the media of choice for most on-line applications. Optical media (CDs, DVDs) are popular in the consumer electronic space. Flash based SSDs are popular for those workloads that exhibit random IOs. Disks are being used in laptops, desktops and storage servers (SANs, NAS, DAS). Tape based WORM media and content addressable based disk storage are providing WORM media capabilities in tape and disk technologies, and thus, these technologies can be used to also store compliance/regulatory data.

(Abridged)

Recommended Reading

150
155
1. Anderson D., Dykes J., and Riedel E. More than an interface- SCSI versus ATA. In Proc. Second Annual Conf. on FAST. San Francisco, CA, 2003.
2. Toigo J. Holy Grail of Network Storage Management. Prentice Hall, Englewood Cliffs, NJ, 2003.
3. Voruganti K., Menon J., and Gopisetty S. Land below a DBMS. SIGMOD Rec., 33(1):64–70, 2004.

Answer the following questions:

- 1) Name some of the persistent storage devices mentioned in the text.
- 2) What are the advantages and disadvantages of tapes?
- 3) Where are WORM tapes useful?
- 4) Name the two dominant tape technologies that are currently on the market.
- 5) What is the most widely used form of persistent storage medium?
- 6) What are the protocols which are currently in use for accessing disks?
- 7) What spin-down modes are used to save on power consumption?
- 8) What are the characteristics of some of the optical storage media?
- 9) What is flash technology?
- 10) Where is flash memory used?

Mark the following statements as *true* or *false*:

- 1) Portable and inexpensive to purchase, tapes are often used for backing up or archiving data.
- 2) Flash is memory technology that has volatile characteristics.
- 3) WORM disks can be written only once.
- 4) NAND has significantly lower storage capacity than NOR.
- 5) MLC is used in solid state drives (SSDs) and SLC is used in consumer appliances like cameras, media players, cell phones, etc.
- 6) An SSD (solid-state drive or solid-state disk) is a storage device that stores persistent data on solid-state flash memory.
- 7) In hybrid disks the flash memory handles the data most frequently written to or retrieved from storage.

Match the following terms with their definitions:

- 1) A library
 - 2) A disk cache
 - 3) Fibre channel
 - 4) A storage area network (SAN)
 - 5) DLT (digital linear tape)
 - 6) A jukebox
-
- a) A mechanism for improving the time it takes to read from or write to a hard disk
 - b) A unit in which optical disk drives are mounted
 - c) A high-speed special-purpose network (or sub-network) that interconnects different kinds of data storage devices
 - d) A collection of physical storage media such as tapes or disks and a way to access them
 - e) A form of magnetic tape and drive system used for computer data storage and archiving
 - f) A type of high speed interconnection

Vocabulary

access time ['æk.ses][taim] – přístupová doba
bursty traffic [bɜːsti] [træfik] – shlukový provoz, shluky dat
cache [kæʃ] – skryš, úkryt
competing [kəm'pi:tɪŋ] – konkurenční
data compliance ['deɪ.tə] [kəm'plai.ənts] ⑤ [-tə] [-] – ochrana dat před přepsáním
wear leveling [weər] ['lev.əlɪŋ] ⑤ [wer] [-] – strategie zápisu a mazání (vyrovnání opotřebení)
focus on st ['fəʊ.kəs] ⑤ ['fou-] – zaměření
format agnostic ['fɔː.mæt] [æɡ'nɒs.tɪk] ⑤ ['fɔ:r-] [-'nɑː.stɪk] – přehrávající jakýkoli formát
gadget ['gædʒ.ɪt] – přístroj, zařízení
goal [ɡəʊl] ⑤ [ɡou] – cíl
high definition [haɪ] [ˌdef.ɪ'nɪʃ.ən] – vysoké rozlišení
idle ['aɪ.dl] – nečinný
life [laɪf] – životnost
marked [mɑːkt] ⑤ [mɑːrkt] – výrazný
non-volatile [non'vɒl.ə.taɪl] ⑤ [non' vaː.lə.tʃəl] – stálý, stabilní
numerous ['njuː.mə.rəs] ⑤ ['nuː-] – početný, četný
persistent [pə'sɪs.tənt] ⑤ [pə-] – stálý, trvalý
read time [riːd] [taɪm] – přístupová doba pro čtení
solid state disk (SSD) ['sɒl.ɪd] [steɪt] [dɪsk] ⑤ ['sɑː.lɪd] [-] [-] – pevný disk na bázi paměťových čipů
standpoint ['stænd.pɔɪnt] – stanovisko, hledisko
to access st ['æk.ses] – dostat se k, mít přístup k
to allocate ['æl.ə.keɪt] – přidělit
to alter ['ɒl.tə] ⑤ ['ɑːl.tʃə] – změnit
to compete [kəm'pi:t] – soutěžit
to determine [dɪ'tɜː.mɪn] ⑤ [-'tɜː-] – určovat, udávat
to emerge [ɪ'mɜːdʒ] ⑤ [-'mɜːdʒ] – objevit se
to handle ['hændl] – zvládnout, vypořádat se
to package ['pæk.ɪdʒ] – zabalit
to spin down [spɪn] [daʊn] – zpomalit otáčení
to spin up [spɪn] [ʌp] – zrychlit otáčení
to vary ['veəri] ⑤ ['ver.i] – lišit se, různit se
to warrant ['wɒr.ənt] ⑤ ['wɔːr-] – zaručit
trade-off ['treɪd.ɒf] ⑤ [-ɑːf] – kompromis

via [vaɪə] ⑤ ['viː.ə] – přes, prostřednictvím
wise (*suffix*) [waɪz] – pokud jde o, hovoříme-li o
write time [raɪt] [taɪm] – přístupová doba zápisu

Storage Protection

KENICHI WADA
Hitachi Limited, Tokyo, Japan

5

Definition

Storage protection is a kind of data protection for data stored in a storage system. The stored data can be lost or becomes inaccessible due to, mainly, a failure in storage component hardware (such as a hard disk drive or controller), a disastrous event, an operator's mistake, or intentional alteration or erasure of the data. Storage protection provides the underlying foundation for high availability and disaster recovery.

Historical Background

In 1956, IBM shipped the first commercial storage that had a hard disk drive. To protect data from bit errors on disk platters, the hard disk drive commonly uses cyclic redundancy check (CRC) and an error-correcting code (ECC).

CRC and ECC cannot protect data from a whole disk failure in which an entire disk becomes inaccessible (for example, because of a disk head crash). The IBM 3990, which was shipped in the 1980s, had the replication functionality in which two identical copies of data were maintained on separate media. This approach protected data from this kind of failure. Replication functionality can be implemented in many other layers of the computer system. Most DBMS support database replication. Some file systems and Logical Volume Managers have file or volume replication functionality. Further, many storage systems and storage virtualization appliances support volume replication functionality.

RAID (Redundant Array of Inexpensive Disks) is another technology for protecting data from whole disk failure. D. Patterson et al. published a paper "A Case for Redundant Arrays of Inexpensive Disks (RAID)" in June 1988 at the SIGMOD conference [6]. This paper introduced a five level data protection scheme. The term RAID was adopted from this paper, but currently RAID is an acronym for Redundant Arrays of Independent Disks. It is noted that the patent covering RAID level 5 technology was issued in 1978 [5].

RAID level 1 is a kind of replication. RAID level 2 to 5 can reduce the capacity required to protect data against disk drive failure than replication, but it is limited to protect disk drive failure. Replication, on the other hand, can be used to protect databases, file systems and logical volume. Further replication can be used for disaster recovery, if data are replicated remotely.

Foundations

Hard disk drives commonly use Reed-Solomon code [7] to correct bit errors. Data in hard disk drives is usually stored in fixed length blocks. Controllers in hard disk drives calculate ECC for each block and record it associated with the original data. When data are read the controller checks data integrity using ECC. CRC can be used with ECC for detecting bit errors and/or reducing the possibility of correction error.

Most DBMS support database replication with master/slave relation between the original and the replica. The master process updates and transfers it to the slave. This type of replication can provide high availability to the client of the DBMS in case of storage system failures as well as server failures. Another type of database replication is multi-master, which is mostly used to provide high performance parallel processing. Both types can be either synchronous or asynchronous replication. In synchronous replication, updates made in original are guaranteed in the replica, note there may be some delay in asynchronous replication.

Volume replication by storage system is also widely accepted as data protection. There are synchronous and asynchronous replications, the same as database replication. Asynchronous volume replication is often used for long distance remote replication. It may prevent performance degradation caused by replication delay, but could cause some data loss in case of recovery. Synchronous replication, on the other hand, may provide no data loss recovery, but may cause performance degradation due to replication delay. Volume replication is also used within a local data center for online backup. Backup servers use replica volume for backup when original volume is online. To support this, a storage system can pause update delegation from original to replica volume.

RAID (Redundant Array of Independent Disks) is a set of disks from one or more commonly accessible disk subsystems, combined with a body of control software, in which part of the physical storage capacity is used to store redundant information about user data stored on the remainder of the storage capacity. The term RAID refers to a group of storage schemes that divide and replicate data among multiple disks, to enhance the availability of data at desired cost and performance levels. A number of standard schemes have evolved which are referred to as levels. Originally, five RAID levels were introduced [6], but many more variations have evolved. Currently, there are several sublevels as well as many non-standard levels. There are trade-offs among RAID levels in terms of performance, cost and reliability.

Key Applications

- 115 Storage protection is essential to achieve business continuity and legal compliance with adequate performance, cost, and reliability.

(Abridged)

120

Recommended Reading

1. ANSI. NFPA1600 Standard on Disaster/Emergency Management and Business Continuity Programs.
- 125 2. BSI. BS25999; Business Continuity Management.
3. Houghton A. Error Coding for Engineers. Kluwer Academic Publications, Hingham, MA, 2001.
4. Keeton K., Santos C., Beyer D., Chase J., and Wilkes J. Designing for disasters. In Proc. 3rd USENIX Conf. on File and Storage Technologies, 2004.
- 130 5. Ouchi N.K. (IBM Corporation). System for recovering data stored in failed memory unit. US Patent 4,092,732, 1978.
- 135 6. Patterson D., Gibson G., and Katz R. A case for redundant arrays of inexpensive disks (RAID). In 1988.
7. Sweeney P. Error Control Coding From Theory to Practice. Wiley, New York, 2002.
- 140 8. <http://www.sec.gov/>

Answer the following questions:

- 1) How can data loss occur?
- 2) What are CRC and ECC?
- 3) What does RAID stand for?
- 4) How many RAID levels were originally introduced?
- 5) What is Reed-Solomon code used for?
- 6) Besides disaster recovery, what is storage protection good for?
- 7) What is the basic difference between RAID 1 and the other RAID levels?

Match the following terms with their definitions:

- 1) synchronous replication
 - 2) asynchronous replication
 - 3) RAID level
 - 4) data integrity check
-
- a) in this approach, changes in the original can take some time to reflect in the replica
 - b) a way to determine whether data is corrupted
 - c) in this approach, changes in the original are immediately reflected in the replica
 - d) a specific strategy of distributing data over multiple disks

Mark the following statements as *true* or *false*:

- 1) When data is read, the controller checks data integrity using CRC.
- 2) Replication can also be used to protect databases and file systems.
- 3) Remote replication often employs the asynchronous volume replication approach.

Vocabulary

appliance [ə'plai.ənt s] – přístroj
array [ə'rei] – pole, sada
compliance [kəm'plai.ənt s] – shoda, dodržení
degradation [,deg.rə'dei.ʃən] – pokles, zhoršení
disastrous [di'zɑ:.strəs] ^{US} [-'zæs.trəs] –
katastrofální, neblahý
disk platter [disk] ['plæt.ər] ^{US} [-] ['plæt.ə] –
disková plotna
erasure [i'rei.ʒər] ^{US} [-zə] – smazání
failure ['feɪ.ljər] ^{US} [ljə] – selhání
inaccessible [ɪn.ək'ses.i.bl] – nepřístupný,
nedostupný
integrity [ɪn'teg.rə.ti] ^{US} [-tɪ] – celistvost,
neporušenost
intentional [ɪn'ten.ʃən.əl] – záměrný
loss [lɒs] ^{US} [lɑ:s] – ztráta
recovery [ri'kʌv.ər.i] ^{US} [-ə-] – obnova
redundancy [ri'dʌn.dən.t.si] – nadbytečnost
replica ['rep.li.kə] – kopie, duplikát
scheme [ski:m] – schéma, soustava
to evolve [ɪ'vɒlv] ^{US} [-'vɑ:lɪv/] – vyvinout se
volume ['vɒl.ju:m] ^{US} ['vɑ:l-] – objem; svazek

Phrases

due to [dju:] ^{US} [du:] – kvůli; způsobený (čím)

Video

YING LI
IBM T.J. Watson Research Center, Hawthorne,
5 NY, USA

Definition

10 Video, which means “I see” in Latin, is an electronic
representation of a sequence of images or frames, put
together to simulate motion and interactivity. From the
producer’s perspective, a video delivers information
created from the recording of real events to be
15 processed simultaneously by a viewer’s eyes and ears.
For most of time, a video also contains other forms of
media such as audio.

Video is also referred to as a storage format for moving
pictures as compared to image, audio, graphics and
20 animation.

Historical Background

Video technology was first developed for television
systems, but it has been further developed in many
25 formats to allow for consumer video recordings.
Generally speaking, there are two main types of video:
analog video and digital video. Analog videos are
usually recorded as PAL (Phase Alternating Line) or
NTSC (National Television System Committee)
30 electric signals following the VHS (Video Home
System) standard and stored in magnetic tapes. Digital
videos, on the contrary, are usually captured by digital
cameras and stored in digital video formats such as
DVD (Digital Versatile Disc), QuickTime and MPEG-
35 4 (Moving Picture Experts Group).

Launched in September 1976, VHS became a standard
format for consumer recording and viewing by the
1990s. Since then, it has dominated both home and
commercial video markets. In March 1997, the DVD
40 format was introduced to American consumers, which
gradually pulled consumers away from VHS in the
following years due to its much better quality. In June
2003, the DVD’s market share exceeded that of the
VHS for the first time. Since then, it has been steadily
45 expanding its consumer market, and by July 2006,
most major film studios have stopped releasing new
movie titles in VHS format, opting for DVD-only
releases. Now, VHS is gradually disappearing from
both rental and retail stores, and DVD has dominated
50 the whole commercial market. Nevertheless, VHS is
still popular for home recording of television
programs, due to the large installed base and the lower
cost of VHS recorders and tape.

For the last few decades, as video technology quickly
55 advances and the cost of storage devices rapidly
decreases, digital videos have become widely available

in diverse application areas such as medicine, remote
sensing, entertainment, education and online
information services. This has thus led to very active
60 research in various video-related areas.

Foundations

The last three decades have witnessed a significant
65 amount of research efforts on various aspects of video
technologies. Roughly speaking, they fall into the
following three general categories: video
representation, video content analysis, and video
application. Specifically, video representation deals
70 with the way a video is represented, in another word,
the file format. Video content analysis, on the other
hand, aims to automatically structure and ultimately
understand the video by analyzing its underlying
content. Due to the difficult nature of this problem,
75 such process usually involves the analysis of multiple
media modalities including visual, audio and text
information. Finally, video application applies what it
has learned from the analysis engine, and facilitates
various types of content access including video
80 browsing, summarization and retrieval. A brief
discussion on each of these three research domains is
given below.

Video Representation

85 A video sequence with accompanying sound track can
occupy a vast amount of storage space when
represented in digital format. As estimated in [6], a
1-min video clip could possibly occupy up to 448 MB.
90 Consequently, compression has been playing an
important role in modern schemes for video
representation.

A wide variety of methods have been proposed to
compress the video stream. Nevertheless, almost all of
95 them build their approaches upon the fact that video
data contains both spatial and temporal redundancy.
Specifically, to reduce the spatial redundancy, an intra-
frame compression is applied which registers
differences between parts of a single frame. Such a
100 task is more closely related to image compression.
Likewise, to reduce the temporal redundancy, an inter-
frame compression is exploited which registers
differences between neighboring frames. This involves
discrete Cosine transform (DCT), motion
105 compensation and other techniques. Some popular
video compression mechanisms include H.261, H.263,
H.264, MPEG-1, MPEG-2, MPEG-4 and MJPEG
(Motion-Joint Photographic Experts Group).
Specifically, H.261 is a 1990 ITU-T
110 (Telecommunication Standardization Sector of
International Telecommunication Union) video coding
standard originally designed for transmission over

ISDN lines. Later on, H.263 and H.264, which provide more capabilities and mainly target at video-conferencing applications, were standardized in 1995 and 2003, respectively. In 1998, the Moving Picture Experts Group (MPEG) was formed to establish an international standard for the coded representation of moving pictures and associated audio on digital storage media. Currently, there have been three established MPEG standards from this effort: MPEG-1, MPEG-2, and MPEG-4. Each of them targets at different commercial applications. For instance, MPEG-1 is usually used as the Video CD (VCD) format, MPEG-2 for High Definition Television (HDTV), and MPEG-4 for streaming video applications. Finally, to facilitate mobile appliances such as digital cameras, MJPEG was developed in the 1990s which uses intra-frame coding technology that is very similar to those used in MPEG-1 or MPEG-2. However, it does not use inter-frame prediction, which on the one hand, results in a loss of compression capability, yet on the other hand, it makes the degree of compression capability independent of the amount of motion in the scene. Moreover, it also eases video editing as simple editing can now be performed at any frame.

Video Content Analysis

Video is a type of rich media as it often consists of other media types such as audio and text. Consequently, research on video content analysis can be grouped into three classes: visual content analysis, audio content analysis, and audiovisual content analysis. A general goal of video content analysis is to extract the underlying video structure so as to facilitate convenient and nonlinear content access. Yet a more aggressive goal is to automatically understand video semantics so as to support applications such as video summarization and retrieval that require an in-depth understanding of the video content.

Video Application

Besides the large amount of research efforts on video content analysis, there are also many attentions on studying various video applications. After all, making the bulky and unstructured video content convenient and efficient to access, present, share, search and deliver is the ultimate goal of the entire research community in this area.

(Abridged)

165 Recommended Reading

1. Cheung S. and Zakhor A. Efficient video similarity

- measurement with video signature. *IEEE Trans. Circ. Syst. Video Tech.*, 13(1):59–74, 2003.
- 170 2. Li Y. and Dorai C. SVM-based audio classification for instructional video analysis. In *Proc. IEEE Int. Conf. on Acoustics, Speech and Signal Processing*, 2004.
3. Li Y. and Kuo C.-C. *Video Content Analysis Using Multimodal Information: for Movie Content Extraction, Indexing and Representation*. Kluwer, MA, USA, 2003.
- 175 4. Li Y., Narayanan S., and Kuo C.-C. Content-based movie analysis and indexing based on audiovisual cues. *IEEE Trans. Circ. Syst. Video Tech.*, 14(8):1073–1085, 2004.
- 180 5. Mahmood T.S. and Srinivasan S. Detecting topical events in digital video. In *Proc. 8th ACM Int. Conf. on Multimedia*, 2000, pp. 85–94.
- 185 6. Mitchell J., Pennebaker W., Fogg C., and LeGall D. *MPEG Video Compression Standard*. Chapman & Hall, New York, NY, USA, 1992.
7. MPEG Requirements Group, *MPEG-7 Applications Document v.8, ISO/MPEG N2860*, MPEG Vancouver Meeting, July 1999.
- 190 8. MPEG Requirements Group, *MPEG-7 Context, Objectives and Technical Roadmap, ISO/MPEG N2861*, MPEG Vancouver Meeting, July 1999.
9. MPEG Requirements Group, *MPEG-7 Requirements Document V.15, ISO/MPEG N4317*, MPEG Sydney Meeting, July 2001.
- 195 10. Nock H., Adams W., Iyengar G., Lin C., Naphade M., Neti C., Tseng B., and Smith J. User-trainable video annotation using multimodal cues. In *Proc. 26th Annu. Int. ACM SIGIR Conf. on Research and Development in Information Retrieval*, 2003, pp. 403–404.
- 200 11. Oh J. and Hua K. Efficient and cost-effective techniques for browsing and indexing large video databases. In *Proc. ACM SIGMOD Int. Conf. on Management of Data*, 2000, pp. 415–426.
- 205 12. Pfeiffer S., Lienhart R., Fischer S., and Effelsberg W. Abstracting digital movies automatically. *J. Vis. Comm. Image Represent.*, 7(4):345–353, 1996.
- 210 13. Yan R., Hauptmann A., and Jin R. Negative pseudo-relevance feedback in content-based video retrieval. In *Proc. 11th ACM Int. Conf. on Multimedia*, 2003, pp. 343–346.
- 215 14. Yeung M., Yeo B., and Liu B. Extracting story units from long programs for video browsing and navigation. In 1996, pp. 296–305.
15. Zhang T. and Kuo C.-C. Audio content analysis for on-line audiovisual data segmentation. *IEEE Trans. Speech Audio Process.*, 9(4):441–457, 2001.
- 220 16. Zheng W., Li J., Si Z., Lin F., and Zhang B. and Using highlevel semantic features in video retrieval. In *Image and Video Retrieval*. Springer, Berlin Heidelberg, New York, 2006, pp. 370–379.

Answer the following questions:

- 1) What is *video*?
- 2) What is the difference between *analog* and *digital* video?
- 3) What makes VHS still popular for home recording?
- 4) What is meant by *remote sensing*?
- 5) What do the terms *video representation*, *video content analysis*, and *video application* refer to?
- 6) What do most compression formats build on?
- 7) What is MPEG-1?
- 8) What does the term “*rich media*” refer to?

Match the following terms and their definitions:

- 1) video representation
 - 2) video content analysis
 - 3) intra-frame compression
 - 4) inter-frame compression
 - 5) likewise
-
- a) a way of reducing spatial redundancy
 - b) deals with the file format
 - c) a way to reduce temporal redundancy
 - d) involves structuring the video
 - e) means the same as “similarly”

Mark the following statements as *true* or *false*:

- 1) Video was first developed for home use.
- 2) H.261 is a video coding standard originally designed for transmission over ISDN lines.
- 3) MPEG-4 is used as a high definition television standard.
- 4) MJPEG has been designed for use in mobile appliances.
- 5) MJPEG has nothing in common with MPEG-1 and MPEG-2 formats
- 6) A general goal of video content analysis is to facilitate convenient and linear content access.

Vocabulary

analysis [ə'næl.ə.sɪs] – analýza, pl. analyses
appliance [ə'plai.ənts] – zařízení
audio ['ɔ:di.əʊ] ^{US} ['a:di.əʊ] – audio, zvuk
audiovisual [ɔ:di.əʊ'vɪz.u.əl] ^{US} [a:di.əʊ-]
– audiovizuální
capability [ˌkeɪ.pə'bɪl.ɪ.ti] ^{US} [-ə.ti] – schopnost
compression [kəm'preʃ.ən] – komprese
consumer [kən'sju:z.mə] ^{US} [-'su:z.mə] – spotřebitel
discrete [dɪ'skri:t] **cosine** ['kəʊ.saɪn] ^{US} ['kou-]
transform [træns'fɔ:m] ^{US} [-'fɔ:rm] – diskrétní kosinová transformace
diverse [daɪ'vɜ:s] ^{US} [dɪ'vɜ:s] – rozdílný
domain [dəʊ'meɪn] ^{US} [dou-] – doména
due to st [dju:] ^{US} [du:] – kvůli něčemu
format ['fɔ:mæt] ^{US} ['fɔ:r-] – formát
frame [freɪm] – rám, rámeček
linear ['lɪn.i.ə] ^{US} [-ə] – lineární
market share ['mɑ:kɪt] [ʃeə] ^{US} ['mɑ:r-] [ʃer]
– podíl na trhu
modality – modalita
motion ['məʊ.ʃən] ^{US} ['mou-] – pohyb
nevertheless [ˌnev.ə.ðə'les] ^{US} [-ə-] – nicméně
non-linear [non'lɪn.i.ə] – nelineární
present ['prez.ənt] – současný (compare the pronunciation with that of verb *to present* [prɪ'zent])
research [rɪ'sɜ:tʃ] ^{US} ['ri:z.ɜ:tʃ] – výzkum
retrieval [rɪ'tri:vəl] – vyhledávání, vyzvedávání
sequence ['si:kwənts] – sekvence
spatial ['speɪ.ʃəl] **and temporal** ['tem.pərə] ^{US} [-pərə]
redundancy [rɪ'dʌn.dənt.sɪ] – prostorová a časová nadbytečnost (redundance)
steady ['sted.i] – stálý
thus [ðʌs] – tak, a tak
to aim to do st [eɪm] – snažit se něco dělat, být zaměřen na děláni něčeho


to browse st [braʊz] – listovat, procházet něčím
to capture st ['kæp.tʃə] ^{US} [-tʃə] – zachytit něco
to deal with st [diəl] – zabývat se něčím
to develop [dɪ'vel.əp] – vyvinout, vyvíjet
to disappear [ˌdɪs.ə'piə] ^{US} [-'pɪr] – zmizet, mizet
to dominate ['dɒm.ɪ.neɪt] ^{US} ['dɑ:mə-] – dominovat
to ease st [i:z] – udělat něco jednodušší, zjednodušit
to exceed st [ɪk'si:d] – přesahovat, překročit něco
to extract st [ɪk'strækt] – extrahovat, vytáhnout něco
to facilitate st [fə'sɪl.ɪ.teɪt] – zjednodušit něco, umožnit něco
to fall into st [fɔ:l] ^{US} [fa:l] – spadat do něčeho
to involve st [ɪn'vɒlv] ^{US} [-'vɑ:lv] – zahrnovat něco
to launch st [lɔ:ntʃ] ^{US} [la:ntʃ] – vypustit něco, vydat něco
to occupy ['ɒk.ju.paɪ] ^{US} ['ɑ:kju-] – zabírat, okupovat
to opt for st [ɒpt] ^{US} [ɑ:pt] – rozhodnout se pro něco, zvolit něco
to present st [prɪ'zent] – prezentovat něco
to refer to st [rɪ'fɜ:] – odkazovat na něco, označovat něco
to release st [rɪ'li:s] – vypustit, vydat něco
to simulate ['sɪm.ju.leɪt] ^{US} [-tɪd] – simulovat, napodobovat
to witness st ['wɪt.nəs] – být něčemu svědkem
ultimate [ˌʌl.tɪ.mət] ^{US} [-tɪə-] – konečný, nejzazší
underlying [ˌʌn.də'laɪ.ɪŋ] ^{US} [-də-] – základní

Phrases

consequently ['kɒn.t.sɪ.kwənt.li] ^{US} ['ka:nt-] – následně
generally speaking – obecně řečeno

independent of st [ˌɪn.dɪ'pen.dənt] – nezávislý na něčem

likewise ['laɪk.waɪz] – podobně

on the contrary ['kɒn.trə.ri]  ['kɑːn.tre-] – naopak

on the one hand ..., on the other hand ... – na jedné straně ..., na druhé straně ...

to play an important role in st – hrát důležitou úlohu v něčem